HIERARCHICAL LIKELIHOOD INFERENCE ON CLUSTERED COMPETING RISKS DATA

by

Nicholas J. Christian

B.A. in Mathematics, University of North Carolina - Asheville, 2005

M.S. in Statistics, University of Texas, 2008

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Nicholas J. Christian

It was defended on

July 14, 2011

and approved by

Stewart Anderson, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Abdus S. Wahed, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

(Joyce) Chung-Chou Ho Chang, PhD
Associate Professor
School of Medicine
University of Pittsburgh

Dissertation Advisor
Jong-Hyeon Jeong, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

HIERARCHICAL LIKELIHOOD INFERENCE ON CLUSTERED COMPETING RISKS DATA

Nicholas J. Christian, PhD

University of Pittsburgh, 2011

Frailties models, an extension of the proportional hazards model, are used to model clustered survival data. In some situations there may be competing risks within a cluster. When this happens the basic frailty model is no longer appropriate. Depending on the purpose of the analysis, either the cause-specific hazard frailty model or the subhazard frailty model needs to be used. In this work, hierarchical likelihood (h-likelihood) methods are extended to provide a new method for fitting both types of competing risks frailty models. Methods for model selection as well as testing for covariate and clustering effects are discussed. Simulations show that in cases with little information, the h-likelihood method can perform better than the penalized partial likelihood method for estimating the subhazard frailty model. Additional simulations demonstrate that h-likelihood performs well when estimating the cause-specific hazard frailty model assuming both a univariate and bivariate frailty distribution. A real example from a breast cancer clinical trial is used to demonstrate using h-likelihood to fit both types of competing risks frailty models.

Public health significance: When researchers have clustered survival data and the observations within those clusters can experience multiple types of events the popular proportional hazards model is no longer appropriate and can lead to biased estimates. For the results of a clinical study to be meaningful the estimated effects of treatments and other covariates needs to be accurate. H-likelihood methods are an alternative to existing procedures and can provide less bias and more accurate information which will ultimately lead to better patient care.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The proportional hazards model (Cox, 1972) is used to describe the relationship between a set of explanatory variables and a possibly right-censored time to event outcome. One of this model's assumptions is that the event times are independent. However, there are situations when this assumption is not appropriate. For example, the event times of related individuals like family members are not independent; siblings share common genetic traits as well as a common environment during childhood. Another example is when individuals may experience an event more than once. Repeated event times for an individual can be influenced by the person's occupation, lifestyle, medical history, etc. In both these examples, the event times are correlated and the proportional hazards model is no longer appropriate. Ignoring the correlation can affect the result of the analysis by leading to underestimated covariate effects (Barker and Henderson, 2005, Ha et al., 2001, Henderson and Oman, 1999).

An extension of the proportional hazards model is the shared frailty model, first introduced by Clayton (1978). The shared frailty model accounts for dependent event times by including a latent random effect called a frailty for each cluster which acts multiplicatively on the hazard function. The frailty term represents the unobserved covariates that explain why the event times are correlated. Given the frailty term, the event times within clusters are conditionally independent. Frailties are assumed to be independent and identically distributed random variables; often a lognormal or gamma distribution is used. The term frailty was first introduced by Vaupel et al. (1979) for modeling population heterogeneity using a univariate frailty model.

In some situations competing risks maybe present within clusters; subjects within a cluster may experience more than one type of event where the occurrence of one of these events prevents observing the other events for this subject. Similarly, subjects who experience

repeated events may experience multiple types of events where a separate terminal event precludes the occurrence of any future events. Only the terminal event censors the other event times and not vice versa, this is a semi-competing risks situation (Fine et al., 2001). For instance in cancer studies, patients can experience multiple event types: local, regional, or distance recurrence as well as a new second primary cancer or death, where only the occurrence of the death prevents the other events from being observable.

The basic frailty model assumes that within clusters censoring times are independent of event times. Under competing risks and semi-competing risks this assumption is no longer reasonable. Factors that affect one event type may also influence the probability of other event types. Therefore it is necessary to incorporate competing risks along with clustering effects in order to obtain unbiased estimates.

Huang and Wolfe (2002) present a model that extends the basic frailty model to include competing risks. The model is fitted using the EM algorithm (Dempster et al., 1977), along with Markov chain Monte Carlo (MCMC) simulation. This is a computationally intensive approach that requires MCMC to perform numerical integration in the E-steps. Another drawback of this approach is that the standard errors are not readily available. Instead they need to be computed using Louis' formula (Louis, 1982). In this work, a more general model is introduced for modeling clustered competing risks by assuming a multivariate distribution for the frailty terms, referred to as the cause-specific hazard frailty model.

Recently, Liu and Huang (2008) proposed using Gaussian quadrature to fit the shared frailty as well as the competing risks model proposed by Huang and Wolfe (2002). Their approach is easy to implement and requires less computation time than the EM algorithm. However the method is not a true non-parametric procedure, a piecewise constant hazard is assumed for the baseline hazard.

Another approach for handling clustered competing risks data is given by Katsahian et al. (2006) and Katsahian and Boudreau (2011) who introduces an estimation procedure for modeling the subhazard function (Fine and Gray, 1999) with clustered data, referred to as the subhazard frailty model. The method proposed by Katsahian and Boudreau (2011) relies on the penalized partial likelihood and can lead to biased estimates when there is little information.

In this work, an alternative estimation procedure is proposed using hierarchical likelihood (h-likelihood) (Ha et al., 2001, Ha and Lee, 2003, Lee and Nelder, 1996). Unlike a traditional likelihood function, the h-likelihood incorporates fixed effects as well as random effects into an extended likelihood. As a result, the method does not require numerically intensive methods to perform numerical integration like the EM-algorithm. Instead parameters are estimated by using the Newton-Raphson method to maximize the profile and adjusted profile h-likelihoods assuming a non-parametric baseline hazard. The Newton-Raphson method, a technique that is faster than using the EM algorithm, has a quadratic convergence rate whereas the convergence rate of the EM algorithm is linear (Tanner, 1996). Moreover, standard errors are a result of the estimation procedure; no additional calculations are needed. Thus, the hierarchical likelihood provides an approach that is faster than the EM algorithm and more general than using Gaussian quadrature.

This work is organized as follows. In Chapter 2 methods for dealing with nuisance parameters in the likelihood function are reviewed and the fundamentals of survival analysis and competing risks are discussed. Then Chapter 3 discusses general characteristics of frailty models as well as describes the basic univariate and shared frailty models and extends these models to the cause-specific hazard frailty model and the subhazard frailty model. Next, Chapter 4 derives the h-likelihood estimators for the cause-specific hazard frailty model and the subhazard frailty model. Methods for statistical inference and model selection are also presented. Chapter 5 presents a simulation study that demonstrates the performance of the hierarchical likelihood estimators for clustered data with competing risks. In Chapter 6, the h-likelihood estimation procedure is applied to a breast cancer dataset from a phase III clinical trial; the h-likelihood can be useful when the goal of a study is the effect of a treatment on multiple outcomes when these outcomes are correlated and subject to competing risks. Finally, Chapter 7 discusses the results and describes future areas of research.

2.0 BACKGROUND

The following chapter reviews relevant material from likelihood theory and survival analysis and introduces notation that will be used throughout this work. Section 2.1 reviews techniques for dealing with nuisance parameters in the likelihood function. Particularly, the adjusted profile likelihood which plays a major role within the h-likelihood estimation procedure. Section 2.2 covers the basics of survival analysis and competing risks which are used for building competing risks frailty models. For a complete discussion, see Pawitan (2001) and Severini (2000) for likelihood methods and Kalbfeisch and Prentice (2002) and Klein and Moeschberger (2003) for survival analysis.

2.1 ADJUSTED PROFILE LIKELIHOOD

Let $f(x|\theta)$ denote the joint density function of a random sample $X = (X_1, X_2, \ldots, X_n)$ of size n. The probability of the observed data $x = (x_1, x_2, \ldots, x_n)$ where θ is the true parameter is given by the likelihood function,

$$L(\theta|x) \equiv f(x|\theta). \tag{2.1}$$

An appropriate estimate of θ is the point where the observed data is most likely or the value that maximizes the likelihood function. Under regularity conditions, the maximum likelihood estimate (MLE), $\hat{\theta}$, is asymptotically normal with mean θ and variance $I(\theta)$, where $I(\theta) = E\left(-\frac{\partial^2}{\partial\theta^2}\log L(\theta|x)\right)$ is the Fisher information matrix. From this asymptotic result Wald hypothesis tests and confidence intervals can be used for approximate inference of θ, where $I(\theta)$ can be estimated by, $I(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\log L(\theta|x)\Big|_{\theta=\hat{\theta}}$.

In this section and throughout this work f, is used to denote a general density function where the arguments and context are used to determine the meaning of the function. Moreover, all likelihood functions are conditional on the observed data. To simplify the notation the observed data x will often be suppressed, $L(\theta) = L(\theta|x)$. The log-likelihood function is denoted by $l(\theta) = \log L(\theta)$.

More complex models require more parameters. Often only a subset of the parameters are of interest to the researcher while the remaining parameters are just used to complete the model and can be thought of as nuisance parameters. To simplify estimation and inference it would be worthwhile to eliminate the nuisance parameters and construct a likelihood function only for the parameters of interest.

Let (θ, η) be all the model parameters where θ is the parameter of interest and η is a nuisance parameter; both θ and η may be vectors. One option is to replace the nuisance parameter η with its MLE for each fixed value of θ; the subsequent likelihood is called the profile likelihood. More formally, given the joint likelihood $L(\theta, \eta)$ the profile likelihood of θ is given by

$$L_p(\theta) = \max_{\eta} L(\theta, \eta) \tag{2.2}$$

where the maximization occurs for a fixed value of θ. Given θ, the MLE of η is often a function of θ. So another way to express the profile likelihood is $L_p(\theta) = L(\theta, \hat{\eta}_\theta)$ or $L_p(\theta) = L(\theta, \hat{\eta}(\theta))$, where $\hat{\eta}_\theta$ and $\hat{\eta}(\theta)$ are both used to denote the MLE of η for a fixed value of θ.

The profile likelihood can be used like a standard likelihood function and does reasonably well with estimation and inference if the number of nuisance parameters is small relative to the sample size. However, the profile likelihood is not a proper likelihood function; it is not based on the distribution of observable data. Therefore there is the potential for biased estimates and underestimated standard errors.

An alternative approach to eliminating nuisance parameters is to use the marginal or conditional likelihoods, which are proper likelihoods. The marginal likelihood is formed by finding a statistic T such that the distribution of T only depends on θ; then the marginal

likelihood is the density of T as a function of θ. Formally, suppose there exists a statistic T such that the density function of the data X may be written as,

$$f(x;\theta,\eta) = f(t;\theta)f(x|t;\theta,\eta). \tag{2.3}$$

Then the marginal likelihood based on the marginal distribution of T is,

$$L(\theta;t) \equiv f(t;\theta). \tag{2.4}$$

The conditional likelihood requires finding a statistic S such that the conditional distribution of the data X given $S = s$ only depends on θ. Then the conditional likelihood is the conditional density function of X given $S = s$ as a function of θ. That is, suppose there exists a statistic S such that,

$$f(x;\theta,\eta) = f(x|s;\theta)f(s;\theta,\eta) \tag{2.5}$$

where S is sufficient for fixed θ. It follows that the conditional likelihood function based on the conditional distribution of X given $S = s$ is,

$$L(\theta;x|s) \equiv f(x|s;\theta). \tag{2.6}$$

Two assumptions are made regarding S. First, S is not sufficient in the full model with parameters (θ,η). If it was, by the factorization theorem $f(x;\theta,\eta) = f(s;\theta,\eta)f(x)$ and there would be no conditional density or corresponding conditional likelihood that only depends on θ. The second assumption is that S does not depend on θ.

Both the marginal and conditional likelihoods eliminate the nuisance parameter η. Moreover, they are both proper likelihoods because they are constructed from a density function, that is they are based on the probability of the observed data. Thus the marginal and conditional likelihoods usually correct the bias and variance of the profile likelihood. One disadvantage of using conditional or marginal likelihoods is that all of the available information about θ may not be used since $f(x|t;\theta,\eta)$ in (2.3) and $f(s;\theta,\eta)$ from (2.5) are ignored. These ignored terms may contain useful information about θ that is not considered. Another disadvantage of these models is that it may not be clear how to find the appropriate statistics. Even if statistics exist, the exact form of the densities can be difficult to derive. One solution

is to approximate the marginal or conditional likelihood by modifying the ordinary profile likelihood (Barndorff-Nielsen, 1983). Below is a heuristic derivation of the approximation (Pawitan, 2001).

First recall that under regularity conditions the MLE of θ, denoted $\hat{\theta}$, is approximately normally distributed with expected value θ and variance $I(\hat{\theta})^{-1}$. Then the approximate density of $\hat{\theta}$ is,

$$f(\hat{\theta}) \approx |I(\hat{\theta})/(2\pi)|^{1/2} \exp\left(-\frac{1}{2} I(\hat{\theta})(\hat{\theta}-\theta)^2\right). \tag{2.7}$$

Furthermore, there is also the quadratic approximation,

$$\log \frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{1}{2} I(\hat{\theta})(\hat{\theta}-\theta)^2 \tag{2.8}$$

found by using a second-order Taylor's series expansion of the likelihood about $\hat{\theta}$. Substituting (2.8) into (2.7) gives the likelihood based p-formula,

$$f(\hat{\theta}) \approx |I(\hat{\theta})/(2\pi)|^{1/2} \frac{L(\theta)}{L(\hat{\theta})}. \tag{2.9}$$

This approximation of the density function of $\hat{\theta}$ turns out to be more accurate than relying on the fact that $\hat{\theta}$ is asymptotically normal (2.7). For the multiparameter case, let $(\hat{\theta}, \hat{\eta})$ denote the MLE of (θ, η) and let $I(\hat{\theta}, \hat{\eta})$ be the corresponding observed information. Then the approximate density of $(\hat{\theta}, \hat{\eta})$ using the multivariate p-formula is,

$$f(\hat{\theta}, \hat{\eta}) \approx |I(\hat{\theta}, \hat{\eta})/(2\pi)|^{1/2} \frac{L(\theta, \eta)}{L(\hat{\theta}, \hat{\eta})}. \tag{2.10}$$

From (2.10) the approximate density function of $\hat{\eta}_\theta$ given θ is,

$$f(\hat{\eta}_\theta) \approx |I(\theta, \hat{\eta}_\theta)/(2\pi)|^{1/2} \frac{L(\theta, \eta)}{L(\theta, \hat{\eta}_\theta)}. \tag{2.11}$$

where $L(\theta, \hat{\eta}_\theta)$ is the profile likelihood of θ. It follows from (2.11) that the marginal density of $\hat{\eta}$ is approximated by using the usual change-of-variable formula,

$$\begin{aligned} f(\hat{\eta}) &= f(\hat{\eta}_\theta) \left| \frac{\partial \hat{\eta}_\theta}{\partial \hat{\eta}} \right| \\ &\approx |I(\theta, \hat{\eta}_\theta)/(2\pi)|^{1/2} \frac{L(\theta, \eta)}{L(\theta, \hat{\eta}_\theta)} \left| \frac{\partial \hat{\eta}_\theta}{\partial \hat{\eta}} \right| \end{aligned} \tag{2.12}$$

Then the conditional density of $\hat{\theta}$ given $\hat{\eta}$ is,

$$\begin{aligned} f(\hat{\theta}|\hat{\eta}) &= \frac{f(\hat{\theta},\hat{\eta})}{f(\hat{\eta})} \\ &\approx \frac{|I(\hat{\theta},\hat{\eta})/(2\pi)|^{1/2}}{|I(\theta,\hat{\eta}_\theta)/(2\pi)|^{1/2}} \frac{L(\theta,\hat{\eta}_\theta)}{L(\hat{\theta},\hat{\eta})} \left|\frac{\partial\hat{\eta}}{\partial\hat{\eta}_\theta}\right| \end{aligned} \tag{2.13}$$

where the p-formula is used on both the numerator and denominator. After ignoring terms in (2.13) that do not depend on the parameters, an approximation to the conditional log-likelihood for θ given $\hat{\eta}$, that is the log-likelihood of θ obtained from $f(\hat{\theta}|\hat{\eta})$ is,

$$l(\theta) = l(\theta,\hat{\eta}_\theta) - \frac{1}{2}\log|I(\theta,\hat{\eta}_\theta)/(2\pi)| + \log\left|\frac{\partial\hat{\eta}}{\partial\hat{\eta}_\theta}\right|. \tag{2.14}$$

Using the marginal density of $\hat{\theta}$, $f(\hat{\theta}) = f(\hat{\theta},\hat{\eta})/f(\hat{\eta}|\hat{\theta})$, a formula identical to (2.14) is found for approximating the marginal likelihood for θ based on $\hat{\theta}$; the likelihood for θ determined by the marginal density of $\hat{\theta}$. Equation (2.14) is referred to as the modified profile likelihood of θ and can be used to approximate either the marginal or conditional likelihood function of θ.

The constant 2π remains in (2.14) so that the formula closely resembles the log of a proper density function. The last term, the Jacobian $|\partial\hat{\eta}/\partial\hat{\eta}_\theta|$ ensures that the modified profile likelihood is invariant with respect to transformations of the nuisance parameter. The Jacobian term can be a very difficult quantity to evaluate. In some cases $\hat{\eta}$ does not depend on θ implying that $\hat{\eta} = \hat{\eta}_\theta$ and $|\partial\hat{\eta}/\partial\hat{\eta}_\theta| = 1$. Generally it is not possible to parameterize the model such that $\hat{\eta} = \hat{\eta}_\theta$. If θ is a scalar, then Cox and Reid (1987) showed that it is possible to set the nuisance parameter η such that $|\partial\hat{\eta}/\partial\hat{\eta}_\theta| \approx 1$ by choosing η such that θ and η are orthogonal. Recall that two parameters are orthogonal if the diagonal elements of the expected information matrix are 0,

$$E\left(\frac{\partial^2}{\partial\theta\partial\eta} l(\theta,\eta)\right) = 0. \tag{2.15}$$

If θ and η are orthogonal parameters then the modified profile likelihood can be approximated by the adjusted profile likelihood defined as,

$$l_A(\theta) = l(\theta,\hat{\eta}_\theta) - \frac{1}{2}\log|I(\theta,\hat{\eta}_\theta)/(2\pi)|. \tag{2.16}$$

The adjusted profile likelihood (2.16) can also be computed using Bayesian methods. Bayesians assume parameters follow a distribution, as a result unwanted parameters can be eliminated by integrating them out.

For scalar parameters, take the quadratic approximation

$$\log \frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{1}{2} I(\hat{\theta})(\hat{\theta} - \theta)^2$$

and integrate both sides with respect to θ then,

$$\begin{aligned} \int L(\theta)\, d\theta &\approx L(\hat{\theta}) \int \exp\left(-\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2\right) d\theta \\ &= L(\hat{\theta}) |I(\hat{\theta})/(2\pi)|^{-1/2}. \end{aligned} \tag{2.17}$$

this is Laplace's integral approximation (Tanner, 1996). If $l(\theta)$ is well approximated by a quadratic then the formula is very accurate. For a two-parameter model $L(\theta, \eta)$ the approximate integrated likelihood for fixed θ is,

$$L_I(\theta) \equiv \int L(\theta, \eta)\, d\eta \approx L(\theta, \hat{\eta}_\theta) |I(\theta, \hat{\eta}_\theta)/(2\pi)|^{-1/2}. \tag{2.18}$$

It follows that the first-order Laplace approximation of the integrated log-likelihood is identical to the adjusted profile likelihood (2.16) when the parameters are orthogonal,

$$l_I(\theta) \approx l(\theta, \hat{\eta}_\theta) - \frac{1}{2} \log |I(\theta, \hat{\eta}_\theta)/(2\pi)|. \tag{2.19}$$

The adjusted profile likelihood is used extensively with hierarchical likelihood methods (Lee and Nelder, 1996, Ha et al., 2001, Lee et al., 2006). In chapter 4 the adjusted profile likelihood is used to approximate the restricted likelihood of the frailty parameter θ.

2.2 COMPETING RISKS

With some studies the outcome of interest is the time until the occurrence of a possibly censored event. More generally, a subject may experience one of m different event types. Or there may be more than one event time for each subject, where the repeated occurrences could either be the same event type or different types of events. In other words, subjects may have multiple failures where each failure is from a different cause.

For example, in an oncology clinical trial studying a new treatment subjects can experience different types of events. They may have a recurrence of the original cancer, development of a new cancer or they may die while in remission. There are also different types of recurrence: local, regional or distance, depending on the site of the recurrence in relation to the site of the initial cancer. Often only the first event is of interest since treatments after the first event mask the effects of the treatment being studied. Thus the occurrence of one of these events precludes us from observing the other event. The m different event types are competing to be observed first and are referred to as competing risks (Prentice et al., 1978, Kalbfeisch and Prentice, 2002). Other times, subjects are followed after the first event and different types of recurrence or new primary cancers are observed until a terminating event, such as death, precludes the occurrence of any future events. Only the terminal event censors the other event times and not vice versa, this is an example of a semi-competing risks situation (Fine et al., 2001).

Competing risks is also referred to as dependent or informative censoring where the occurrence of the competing event censors the event of interest informatively. For example, suppose subjects drop out of a study because of treatment side effects. The potential event time of these subjects is related to experiencing these side effects. Since there is an association between the event of interest and side effects, patients who dropped out of the study are censored informatively; side effects should be considered a competing event. Subjects where not censored randomly but where informatively censored for reasons related to treatment.

Many survival analysis methods assume that event and censoring times are independent. With competing risks data this is assumption may not be reasonable. Factors that affect one event type may also influence the probability of other competing events. Treatment may

lower the probability of recurrence for a subject, but may also increase their probability of death due to treatment toxicity. So it may not be appropriate to treat time of death as an independent censoring time when studying the time to recurrence.

On the other hand, there may be cases when competing risks are independent in which case standard survival analysis methods that assume independent censoring can be used. The problem is that there is not enough information in the data to test the assumption of dependent or independent competing risks (Tsiatis, 1975). Observing only one event type per subject does not provide sufficient information to estimate the association between event types. There is always some degree of informative censoring in a study. Depending on the nature of the study as well as the question being investigated competing risks may or may not be worth accounting for. When necessary there are several ways to deal with dependent competing risks (Moeschberger and Klein, 1995, Zheng and Klein, 1995). One approach is to use quantities that do not make any assumptions about the dependence structure between competing risks.

Using a statistical method that assumes independent competing risks when there is actually an association can lead to bias results and faulty conclusions. A classic example is the complement of the Kaplan-Meier estimator (Kaplan and Meier, 1958) of the survival function, denoted $1 - KM$. It is well known (Pepe and Mori, 1993, Gooley et al., 1999) that this estimator overestimates the cumulative probability of an event of interest, in the presence of competing risks. The reason for the overestimation is how $1 - KM$ handles the other event types. Subjects who experience a competing risk are treated just like subjects who are independently censored for reasons like lost to follow-up. The overestimation is a result of including subjects who failed from a competing risk and are not capable of a future failure with subjects who have not had any events and are still at risk of failure. The $1-KM$ estimator only considers the event of interest and does not account for events from other causes. Thus, $1 - KM$ is interpreted as the probability of an event of interest by time t if the competing events were removed or did not exist. A more appropriate estimator is the cumulative incidence function discussed below.

There are several useful quantities for analyzing competing risks data: cause-specific hazard functions, cumulative incidence functions and subhazard functions. These quantities

are useful because they can be estimated and are identifiable from the observed data without making any assumptions about the joint distribution of the event times.

Before discussing competing risks quantities the following survival notation is quickly introduced; see Klein and Moeschberger (2003) for more details. Let T be a nonnegative random variable from a homogeneous population that denotes the time to an event. Further assume that T is continuous, similar quantities exists when T is discrete. This work assumes that T is subject to right censoring, where all that is known is that a subject has yet to experience an event by a given time.

The survival function is the probability of experiencing the event after time t or, in other words, the probability of surviving until time t,

$$S(t) = P(T > t) = \int_t^\infty f(t)\, dt, \tag{2.20}$$

where $f(t)$ is the probability density function of T. The survival function is a monotone non-increasing function equal to one at zero and approaches zero as time goes to infinity. The survival function can also be calculated using the cumulative distribution function $F(t)$ of T, $S(t) = 1 - F(t)$. Thus, there is a connection between $f(t)$ and $S(t)$,

$$f(t) = \frac{dF(t)}{dt} = \frac{d(1 - S(t))}{dt} = -\frac{dS(t)}{dt}. \tag{2.21}$$

The hazard function or hazard rate is the instantaneous event rate at time t given an individual has reached time t without experiencing the event. The hazard function is defined as,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.22}$$

Unlike the survival function the only restriction on the hazard function is that it be non-negative; hazard rates can be constant, increasing, decreasing, convex or concave. The

hazard function can be expressed in terms of $f(t)$ and $S(t)$,

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t P(T > t)} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t P(T > t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{2.23}$$

Using (2.21) and (2.23) the hazard function can also be written as,

$$\lambda(t) = -\frac{d}{dt} \log(S(t)). \tag{2.24}$$

The cumulative hazard function is,

$$\Lambda(t) = \int_0^t \lambda(u)\, du = -\log(S(t)) \tag{2.25}$$

where the last term comes from integrating both sides of (2.24). Finally, it follows from (2.25) that,

$$S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u)\right)\, du. \tag{2.26}$$

Now back to competing risks, let T_k, $k = 1, 2, \ldots, m$ denote the event time for event type k. Then the observed event time T in the presence of competing risks is the minimum of all of the potential event times $T = \min(T_1, T_2, \ldots, T_m)$. Define the event indicator $\delta_k = 1$ if the type k event occurs first, $T = T_k$, and 0 otherwise.

A useful quantity for modeling competing risks is the cause-specific hazard rate for event k,

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, \delta_k = 1 | T \geq t)}{\Delta t}. \tag{2.27}$$

The function $\lambda_k(t)$ is the rate of experiencing a type k event in the next instance given that the subject has yet to experience any of the competing events. If the event types are

mutually exclusive and cannot occur simultaneously, then the overall hazard rate is the sum of the cause-specific hazard rates,

$$\lambda(t) = \sum_{k=1}^{m} \lambda_k(t) \tag{2.28}$$

and the overall survival rate is $S(t) = \exp(-\int_0^t \sum_{k=1}^m \lambda_k(u)\, du)$. The cumulative hazard rate for event k is $\Lambda_k(t) = -\int_0^t \lambda_k(u)\, du$.

The cumulative incidence function $F_k(t)$ is the probability of the event of interest occurring before time t where an individual is exposed to all event types,

$$F_k(t) = P(T \leq t, \delta_k = 1) = \int_0^t f_k(u)\, du \tag{2.29}$$

where $f_k(t) = \frac{d}{dt} F_k(t)$ is the subdensity function. Notice that the cumulative incidence function is an improper distribution since,

$$\lim_{t\to\infty} F_k(t) = \lim_{t\to\infty} P(T \leq t, \delta_k = 1) = \lim_{t\to\infty} P(T \leq t | \delta_k = 1) P(\delta_k = 1) = P(\delta_k = 1).$$

Thus $F_k(t)$ is also referred to as a subdistribution function.

Since $F(t) = \sum_{k=1}^m F_k(t) \leq 1$, it follows that the probability of one event type will be low if the probability of the other events is high because the sum of cumulative incidence functions cannot exceed one. Therefore, estimators for all event types need to be considered simultaneously in order to give an appropriate interpretation (Korn and Dorey, 1992, Pepe and Mori, 1993).

It may seem worth considering $1 - F_k(t)$. However this quantity has a very awkward interpretation, it is the probability that nothing happens until time t, but when something does happen it needs to be event k. The cumulative incidence function $F_k(t)$ gives a more natural interpretation, the probability of event k by time t (in the presence of competing risks).

The subhazard function is the hazard function of the subdistribution,

$$\begin{aligned} \gamma_k(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, \delta_k = 1 | T \geq t \cup (T \leq t \cap \delta_k = 0))}{\Delta t} \\ &= \frac{f_k(t)}{1 - F_k(t)} \\ &= -\frac{d}{dt} \log(1 - F_k(t)). \end{aligned} \tag{2.30}$$

It follows, that the cumulative incidence function is a function of the cause-specific hazard or the subhazard. Following the steps used to get (2.23) it is easy to show that $f_k(t) = S(t)\lambda_k(t)$. Then,

$$F_k(t) = 1 - \exp\left(-\int_0^t \gamma_k(u)\, du\right) = \int_0^t S(u)\lambda_k(u)\, du \tag{2.31}$$

Differentiating both sides of (2.31) with respect to t and rearranging the terms gives the following relation between the subhazard function and the cause-specific hazard function,

$$\lambda_k(t) = \left(\frac{1 - F_k(t)}{S(t)}\right)\gamma_k(t). \tag{2.32}$$

Lastly, let t_i be the observed event time for the ith subject in a sample of size n and let δ_{ik} be the type k, $k = 1, 2, \ldots, m$ event indicator for subject i when there are m event types such that $\delta_{ik} = 1$ if a type k event occurs at time t_i and $\delta_{ik} = 0$ otherwise. Then the likelihood assuming independent censoring is,

$$L = \prod_{k=1}^{m}\prod_{i=1}^{n} \lambda_k(t_i)^{\delta_{ik}} \exp\left(-\Lambda_{0k}(t_i)\right). \tag{2.33}$$

where Λ_{0k} is the baseline cumulative hazard function.

3.0 FRAILTY MODELS

In some studies, the observed event times may not be independent. For example, siblings share common genetic traits as well as a common environment during childhood. These shared characteristics tend to make the event times of family members correlated. However, different families have different characteristics so event times between families may be independent. Another example is a multi-center clinical trial. A subject's outcome may be influenced by the type of care received at a particular institution. Therefore the event times for subjects from the same center may be correlated while the event times between different centers with different practices may be independent. Repeated outcomes recorded on the same individual over time will also tend to be related. Individuals have unique lifestyles, medical history and genetic traits that can make repeated event times for the same subject dependent. In all of these examples there are clusters of similar observations. The event times within these clusters tend to be correlated because members of a cluster share certain characteristics. Event times between clusters are typically independent because different clusters have different attributes.

When the event times are correlated, the proportional hazards model is no longer appropriate since this model assumes independent event times. Not recognizing the dependence within the data and fitting the usual proportional hazards model can lead to underestimating standard errors and covariate effects (Barker and Henderson, 2005, Ha et al., 2001, Henderson and Oman, 1999).

An extension of the proportional hazards model is the frailty model introduced by Vaupel et al. (1979) and Clayton (1978). The frailty model accounts for dependent event times by incorporating a random effect, called a frailty, that acts multiplicatively on the baseline hazard function of the proportional hazards model (3.1). This random effect represents

unobserved covariates that describe the correlated event times within clusters. For example, when studying siblings the random effect represents the effect of common genetic traits and a common childhood environment which can cause event times between siblings to be dependent. All members of a particular cluster share these unobserved covariates and so every member of a cluster has a common frailty effect. Event times between clusters are assumed to be independent; therefore, frailties are independent random variables. Frailties are also unobservable, since it is not possible to quantify the latent covariates that account for the dependence. When clusters have only one observation, the frailty term adjusts for individual unobserved heterogeneity.

Suppose there are $i = 1, 2, \ldots, n$ clusters where each cluster has $j = 1, 2, \ldots, n_i$ observations, so that the total sample size is $N = \sum_{i=1}^{n} n_i$. The two indices i and j denote a unique observation from the overall sample of size N. Let U_i be an unobserved non-negative random effect or frailty that is shared by each member of cluster i. Given $U_i = u_i$, the hazard function for the jth observation of the ith cluster is,

$$\lambda_{ij}(t|u_i) = u_i \lambda_0(t) \exp(X_{ij}^T \beta) \tag{3.1}$$

where $\lambda_0(t)$ is the baseline hazard function, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a $p \times 1$ vector of fixed parameters and X_{ij} is a $p \times 1$ vector of known covariates for the jth observation of the ith cluster; let X be a $N \times p$ matrix whose ijth row is X_{ij}^T. All covariates are assumed to be fixed and not dependent on time.

The frailties, U_i, are non-negative, independent and identically distributed random variables with a density function having frailty parameter θ (possibly a vector) and are assumed to remain constant over time; Wintrebert et al. (2004) and Yau and McGilchrist (1998) investigate models with time-dependent frailty terms. The basic idea of the frailty term is that groups or individuals who are more frail will have a large value for U_i and be more likely to experience the event earlier, while those who are less frail will have a smaller value for U_i and be likely to experience the event later. Conditional on the frailty term U_i, the event times within cluster i are assumed to be independent. Event times are conditionally independent given the frailty because frailties represent unobserved covariates and event times conditional on covariates are assumed to be independent. In short, the frailty U_i plays two

roles (Wu, 2010): (1) it accounts for the correlation between event times within the same cluster and (2) it accounts for the variation in event times between different clusters. For a homogeneous study population, $U_i = 1$ for all i and (3.1) reduces to the familiar proportional hazards model.

An alternative form of (3.1) treats the frailty more like a random effect from mixed models (McGilchrist, 1993),

$$\lambda_{ij}(t|v_i) = \lambda_0(t) \exp(X_{ij}^T \beta + v_i) \tag{3.2}$$

where $v_i = \log(u_i)$. To help distinguish between models (3.1) and models (3.2) U will be referred to as a frailty and V will be referred to as a random effect. In (3.2), the random effect V_i has more of a symmetric interpretation, where $V_i > 0$ indicates more frail clusters, $V_i < 0$ less frail clusters and $V_i = 0$ indicates no clustering effect. Estimation in chapter 4 is performed using (3.2), since v_i can be any real number whereas u_i is restricted to non-negative values; this restriction may cause problems with convergence.

The frailty model (3.1) is relatively simple. First, it assumes that U_i is constant over time. Thus the frailty term only represent differences between clusters at the start of the study. Second, the model assumes that frailties act proportionally on a common baseline hazard rate shared by every cluster. This does not allow for a lot of uniqueness between clusters since the conditional hazard rate for each cluster follows the same general shape. Nonetheless, the frailty model is still able to capture some of the effects of unobserved covariates providing some understanding of what is going on. Quoting G.E. Box, "All models are wrong, some models are useful."

This chapter begins with a discussion of the general characteristics of frailty distributions and highlights commonly used frailty distributions. Next, several different frailties models are reviewed. First, the simple univariate and shared frailty models are briefly discussed. Then section 3.4 uses correlated frailty models to model cause-specific hazard rates when competing event times are correlated. Lastly, section 3.5 discusses the subhazard frailty model. Both Wienke (2011) and Duchateau and Janssen (2008) are excellent resources for frailty models.

3.1 GENERAL CHARACTERISTICS

3.1.1 Frailty Distributions

Most often either the gamma distribution or lognormal distribution is assumed for the frailty term. The gamma distribution is attractive for mathematical and computational reasons and the lognormal distribution is widely used because of the link between frailty models and generalized mixed models which assumes that random effects are normally distributed. Other commonly used distributions are the inverse Gaussian and positive stable. Both these distributions also share the computational benefits of the gamma distribution. Recall that frailties are non-negative, thus the support of a frailty distribution must be non-negative.

The gamma density function with shape parameter a and scale parameter b is,

$$f(u) = \frac{1}{\Gamma(a)} b^a u^{a-1} e^{-ub} \tag{3.3}$$

for $u \geq 0$ and $a, b > 0$ with $E(U) = a/b$ and $\text{Var}(U) = a/b^2$. This density function is frequently assumed because the Laplace transform has a simple form which makes it easy to calculate the marginal survival and hazard functions (section 3.1.2). Moreover the gamma distribution is very flexible and can take a variety of shapes, Figure 3.1. In order to make sure the frailty model is identifiable the restriction $a = b$ is often made for the gamma distribution. Then $E(U) = 1$ and $\text{Var}(U) = 1/b = \theta$. Thus the gamma density used with frailty models only depends on one parameter θ,

$$f(u) = \frac{1}{\Gamma(1/\theta)} (1/\theta)^{(1/\theta)} u^{1/\theta - 1} e^{-u/\theta} \tag{3.4}$$

where $\theta \geq 0$. It is important to note that there is no biological reason for assuming a gamma distribution, the main advantages of assuming a gamma distribution are computational.

The lognormal frailty distribution is used because of the similarities between mixed models and frailty models. Recall equation (3.2), $\lambda_{ij}(t|v_i) = \lambda_0(t) \exp(X_{ij}^T \beta + v_i)$, following mixed model theory the random effect V_i is assumed to be normally distributed with mean

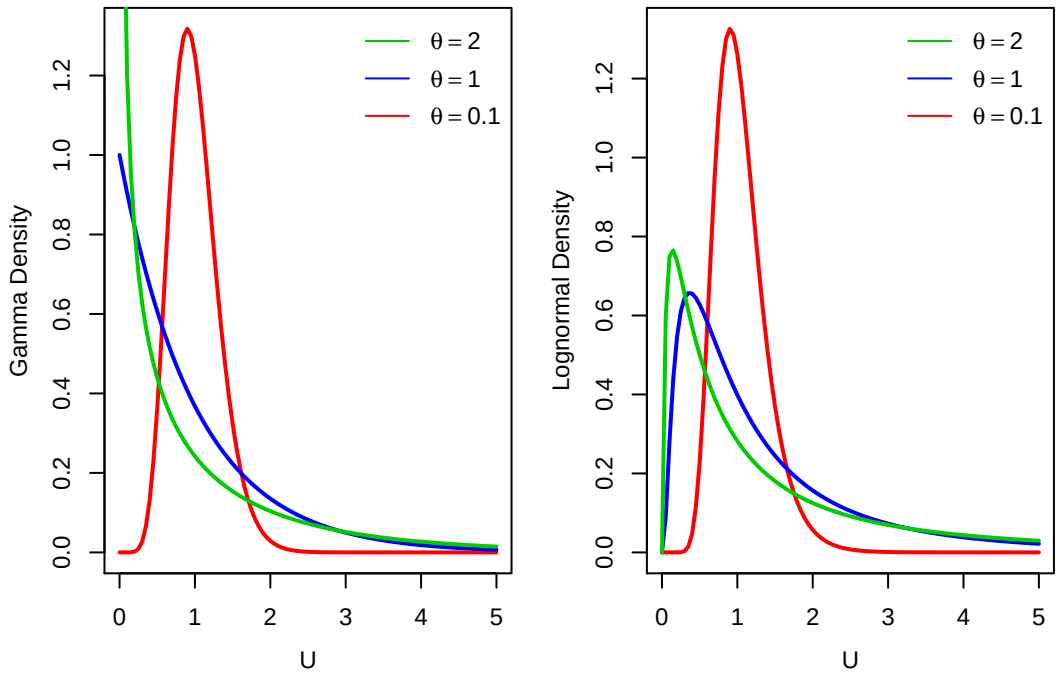

Figure 3.1: Gamma frailty density (left) with mean 1 and variance θ and lognormal frailty density (right) where $V = \log(U)$ has mean 0 and variance θ.

0 and variance θ. Thus the frailty $U = \exp(V)$ is a lognormal random variable with density function,

$$f(u) = \frac{1}{u\sqrt{2\pi\theta}} \exp\left(-\frac{(\log(u))^2}{2\theta}\right) \tag{3.5}$$

where $u \geq 0$ and $\theta \geq 0$. The $E(U) = \exp(\theta/2)$ and $\text{Var}(U) = \exp(2\theta) - \exp(\theta)$. Rather than restrict $E(U) = 1$, to guarantee identifiability, it is more natural to assume that $V \sim N(0, \theta)$. Thus the expected value of a lognormal frailty U will not be one.

A disadvantage of using a lognormal frailty distribution is that there is no closed form solution for the marginal survival function (section 3.1.2). As a result more sophisticated numerical techniques need to be used to numerically evaluate integrals. Despite the computational difficulties, Vaida and Xu (2000) prefer assuming normally distributed random effects. Unlike gamma distributed frailties, normally distributed random effects are symmetric and scale-invariant. Another advantage of lognormal distributions is that for multivariate

frailty problems we can assume a multivariate normal distribution for the random effects V, which will make it easier to model the dependence between frailties, like in section 3.4.

Selecting which frailty distribution to assume depends on the purpose of the analysis as well as available software. If the primary objective of the analysis is the association between members of a cluster then the frailty distribution needs to be carefully considered. Since different frailty distributions lead to noticeably different association structures within clusters (Shih and Louis, 1995). On the other hand, if the interest is on the regression coefficients then selecting a frailty distribution is less important. Simulation studies suggest that misspecifying the frailty distribution has a minimal effect on the estimated regression coefficients (Glidden and Vittinghhoff, 2004, Ha and Lee, 2003). Thus depending on the purpose of the analysis the frailty distribution may or may not have a serious impact.

3.1.2 Frailty Density among Survivors

The frailty density $f(u|\theta)$ with parameter θ describes the frailty in the population at the beginning of the study. For each cluster the frailty effect is assumed to remain constant with respect to time. However, as time goes by the population at risk changes. On average, individuals from more susceptible groups with a higher frailty will have an event earlier and those individuals from less frail groups will have a lower frailty and experience an event later. As a result the distribution of the frailties among survivors changes with time.

Let $f(u|T > t)$ denote the frailty density among the survivors at time t. To illustrate how $f(u|T > t)$ changes with time consider a simple univariate frailty model with one observation per cluster $n_i = 1$ that only consists of a frailty term U_i acting multiplicatively on the baseline hazard function $\lambda_0(t)$ with no covariate information. Then the conditional hazard function for individual i is,

$$\lambda_i(t|u_i) = u_i\lambda_0(t). \tag{3.6}$$

The following results directly extend to clusters of any size. For simplicity, the subscript i will be suppressed for the remainder of this section.

It follows from (3.6) that the survival function for an individual conditional on the frailty $U = u$ is,

$$S(t|u) = \exp\left(-\int_0^t \lambda(s|u)\, ds\right) = \exp\left(-u\Lambda_0(t)\right) \tag{3.7}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$ is the cumulative baseline hazard function. Currently the model is conditional on knowing the unobservable frailties and does not represent what is actually observed. Since the frailties are unobservable it is reasonable to integrate out the frailty random variable and consider the population survival function $S(t)$, the unconditional survival function for an individual randomly drawn from the population being studied. The survival function $S(t)$ is also referred to as the marginal survival function, while $S(t|u)$ is called the conditional survival function. Integrating out the frailty with respect to the frailty distribution gives,

$$\begin{aligned} S(t) &= \int_0^\infty S(t|u) f(u)\, du \\ &= \int_0^\infty \exp\left(-u\Lambda_0(t)\right) f(u)\, du \\ &= E_U\left[\exp\left(-u\Lambda_0(t)\right)\right] \\ &= \mathcal{L}_U\left(\Lambda_0(t)\right) \end{aligned} \tag{3.8}$$

where E_U denotes taking the expectation with respect to the distribution of U and $\mathcal{L}_U$ is the Laplace transform of the frailty density function,

$$\mathcal{L}_U(s) = E_U(e^{-su}) = \int_0^\infty e^{-su} f(u)\, du. \tag{3.9}$$

Laplace transforms play an important role with frailty models. If the Laplace transform of the frailty density has an explicit form then the derivatives of the Laplace transform can be used to obtain general results (Hougaard, 1984). For example, the unconditional density function of T is,

$$f(t) = -\lambda_0(t)\mathcal{L}_U'(\Lambda_0(t)) = -\frac{d}{dt}S(t) \tag{3.10}$$

and the population or marginal hazard function of the event times are,

$$\lambda(t) = -\lambda_0(t)\frac{\mathcal{L}_U'(\Lambda_0(t))}{\mathcal{L}_U(\Lambda_0(t))} = -\frac{d}{dt}\log(S(t)). \tag{3.11}$$

Moreover, the expected value and variance of the frailties can also be expressed as derivatives of the Laplace transform of the frailty density,

$$\begin{aligned} E(U) &= -\mathcal{L}'_U(0) \\ \text{Var}(U) &= \mathcal{L}''_U(0) - (\mathcal{L}'_U(0))^2. \end{aligned}$$

All of these calculations are easy to compute if the Laplace transform has a simple form. Frailty distributions with an explicit Laplace transform like the gamma distribution and inverse Gaussian distribution are often assumed because they simplify parameter estimation. For the lognormal distribution no closed form expression of the Laplace transform exists. As a result most procedures for fitting frailty models with a lognormal frailty distribution require numerical integration methods to calculate the marginal survival function.

Now by Bayes' theorem, the frailty density among the survivors at time t is,

$$f(u|T > t) = \frac{S(t|u)f(u)}{S(t)}. \tag{3.12}$$

Similarly, the frailty density of the individuals who have an event at time t is,

$$f(u|T = t) = \frac{f(t|u)f(u)}{f(t)}. \tag{3.13}$$

Suppose U follows a gamma distribution with mean 1 and variance θ (3.4). Then, the probability of a random individual in the study population surviving until time t is given by

$$S(t) = \mathcal{L}(\Lambda_0(t)) = \left(\frac{1}{1+\theta\Lambda_0(t)}\right)^{\frac{1}{\theta}}. \tag{3.14}$$

It follows from (3.12) that the frailty density for the survivors at time t is,

$$\begin{aligned} f(u|T > t) &= \frac{\exp(-u\Lambda_0(t))\left[u^{1/\theta-1}e^{-u/\theta}/(\theta^{1/\theta}\Gamma(\frac{1}{\theta}))\right]}{(1+\theta\Lambda_0(t))^{-1/\theta}} \\ &= \frac{(\frac{1}{\theta}+\Lambda_0(t))^{1/\theta}}{\Gamma(\frac{1}{\theta})}u^{1/\theta-1}\exp\left(-u[1/\theta+\Lambda_0(t)]\right) \end{aligned} \tag{3.15}$$

a gamma density with shape parameter $1/\theta$ and scale parameter $[1/\theta + \Lambda_0(t)]$. In a similar fashion, $f(u|T = t)$ is also a gamma density with shape parameter $1/\theta+1$ and scale parameter $[1/\theta + \Lambda_0(t)]$.

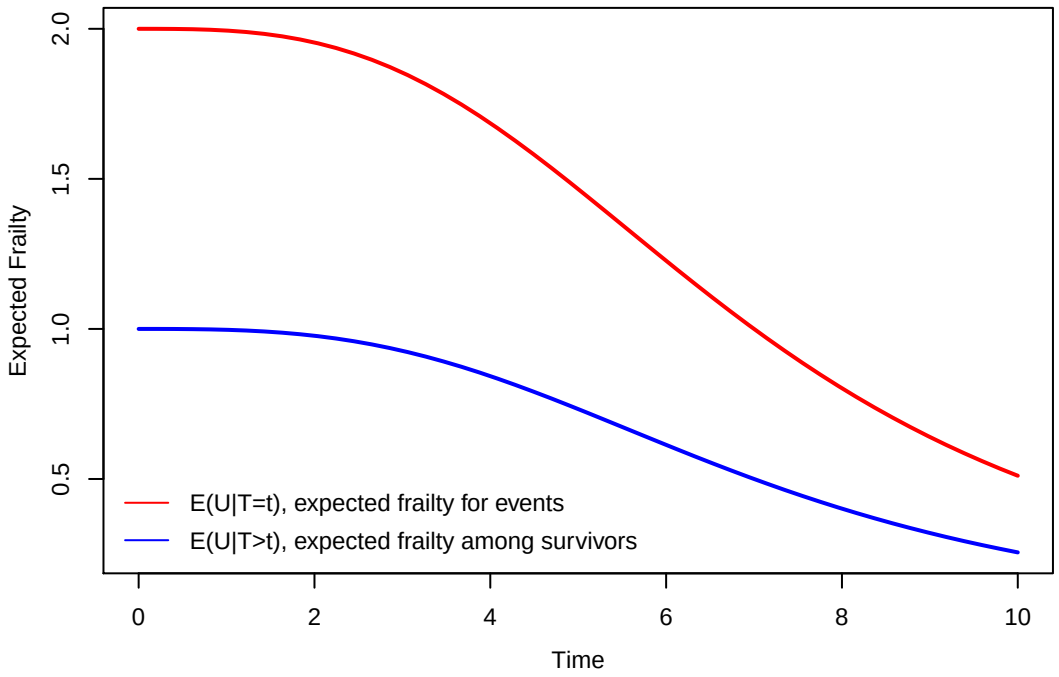

Figure 3.2: Expected frailty for events at time t and among survivors at the same time; where U follows a gamma distribution with mean 1 and variance $\theta = 1$ and $\Lambda_0(t) = (t/7)^3$.

Then the average frailty among the survivors at time t is,

$$E(U|T > t) = \frac{1}{1 + \theta\Lambda_0(t)} \tag{3.16}$$

and the average frailty among individuals who have an event at time t is,

$$E(U|T = t) = \frac{1 + \theta}{1 + \theta\Lambda_0(t)}. \tag{3.17}$$

Notice that both $E(U|T > t)$ and $E(U|T = t)$ decrease over time. The decrease is faster for more heterogeneous populations (large θ) and when the event is not rare (higher cumulative baseline hazard $\Lambda_0(t)$). The individuals who survive the longest will have on average smaller frailties than those who have an event earlier. Furthermore, $E(U|T > t)$ is less than $E(U|T = t)$ for all t. Thus the individuals who have an event at time t have on average a higher frailty than those individuals who survive beyond time t. Figure 3.2 illustrates this relationship assuming a Weibull distribution for T with cumulative baseline hazard $(t/7)^3$ and with $\theta = 1$.

Note that for any frailty distribution, the expected frailty among the survivors decreases with time.

From (3.15) the variance of the frailties among the survivors at time t is,

$$\text{Var}(U|T > t) = \frac{\theta}{(1 + \theta\Lambda_0(t))^2}. \tag{3.18}$$

Likewise, the variance among those who have an event at the same time is,

$$\text{Var}(U|T = t) = \frac{\theta(1 + \theta)}{(1 + \theta\Lambda_0(t))^2}. \tag{3.19}$$

Both (3.18) and (3.19) decrease over time, so it appears that the population is becoming less heterogeneous over time. However, a normalized measure of the variance, such as the coefficient of variation, should be used to describe the heterogeneity of the frailties (Hougaard, 1984). When the frailties follow a gamma distribution with mean 1 and variance θ, the coefficient of variation among survivors at time t is,

$$\frac{\sqrt{\text{Var}(U|T > t)}}{E(U|T > t)} = \sqrt{\theta}. \tag{3.20}$$

Thus the population does not become more homogeneous with time.

Other frailty distributions lead to different conclusions. For example, consider the inverse Gaussian density (Hougaard, 1984),

$$f(u) = \left(\frac{\alpha}{2\pi u^3}\right)^{1/2} \exp\left(-\frac{\alpha}{2u\mu^2}(u - \mu)^2\right) \tag{3.21}$$

with $u \geq 0$ and $\alpha, \mu > 0$; the $E(U) = \mu$ and $\text{Var}(U) = \mu^3/\alpha$. Let $\mu = 1$ so that $\text{Var}(U) = 1/\alpha = \theta =$ then the coefficient of variation for the frailties among the survivors is,

$$\frac{\sqrt{\theta}}{\sqrt{1 + \theta\Lambda_0(t)}}. \tag{3.22}$$

In this case, the coefficient of variation is decreasing as time increases and the population is becoming more homogeneous over time.

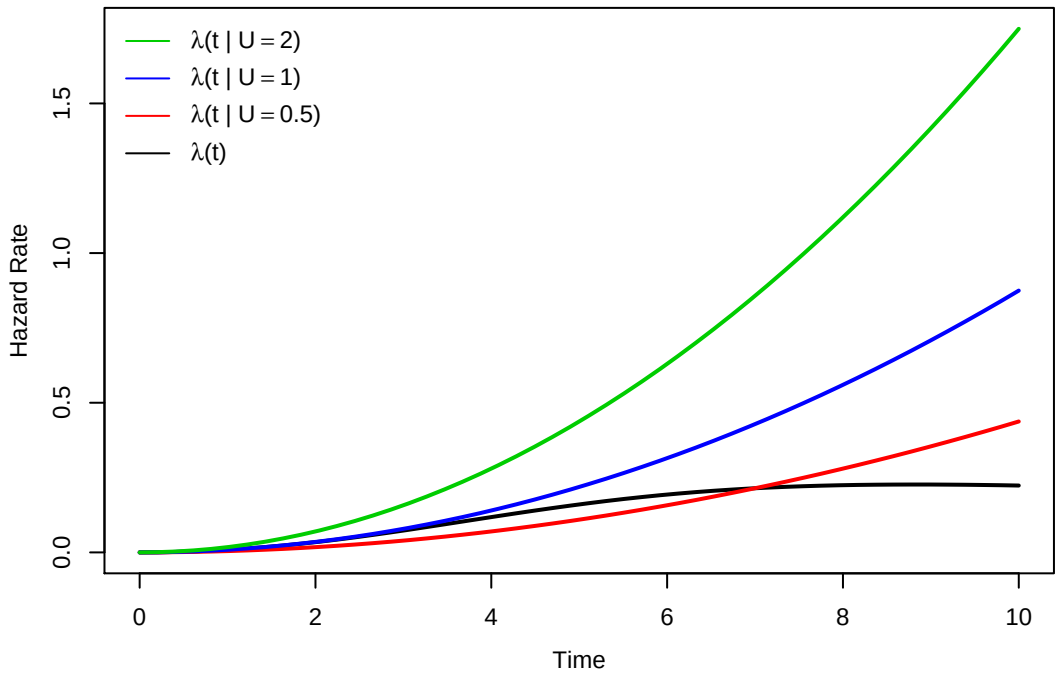

Figure 3.3: Marginal and conditional hazard functions; where U follows a gamma distribution with mean 1 and variance $\theta = 1$ and $\Lambda_0(t) = (t/7)^3$.

3.1.3 Marginal and Conditional Hazard Functions

The marginal or population hazard function, $\lambda(t)$, applies to the entire population, unconditional on the observation's cluster; whereas, the conditional hazards model $\lambda(t|u)$ only applies to observations within a particular cluster. Consider the simple frailty model from section 3.1.2,

$$\lambda(t|u) = u\lambda_0(t).$$

Notice that the baseline hazard function behaves in a similar way for all clusters, if $\lambda_0(t)$ is increasing then the baseline hazard increases over time for every cluster regardless of the frailty term for that cluster. However this is not necessarly true for the population hazard function. If the true model is the conditional hazard function then the derived population hazard function may not behave the same, it might actually decrease when the conditional hazard increases.

Vaupel et al. (1979) shows that the population hazard function can be interpreted as the weighted baseline hazard where the weights are determined by the expected frailty among the survivors,

$$\lambda(t) = \int_0^\infty \lambda(t|u) f(u|T > t)\, du = \lambda_0(t) \int_0^\infty u f(u|T > t)\, du = \lambda_0(t) E(U|T > t). \quad (3.23)$$

Since the expected frailty declines over time, hazard rates for the population will also decrease over time. Thus it is likely that the population hazard rate will not resemble the hazard rate of an individual from that population. Assume a gamma distribution for U then using (3.16) the population hazard rate is,

$$\lambda(t) = \frac{\lambda_0(t)}{1 + \theta \Lambda_0(t)}. \quad (3.24)$$

As in Figure 3.2, assume U follows a gamma distribution with mean 1 and variance $\theta = 1$ and assume a Weibull distribution for T with cumulative baseline hazard rate $\lambda_0(t) = (t/7)^3$. Then, Figure 3.3 is a plot of the population hazard rate and the hazard rate conditional on different frailty values. Notice that conditional on the frailties the hazards increase over time while the population hazard decreases to zero. From Figure 3.3, it is clear that erroneous conclusions can be made if decisions on an individual level are only made based on the population hazard function; Vaupel and Yashin (1985) and Aalen (1994) give additional examples.

3.1.4 Marginal and Conditional Hazard Ratio

Even though the proportional hazards assumption is true for the conditional hazard function $\lambda(t|u)$, this assumption may not be true for the marginal hazard function, $\lambda(t)$. Consider a simple univariate frailty model with $n_i = 1$, where there is just one binary covariate X_i,

$$\lambda_i(t|u_i) = u_i \lambda_0(t) \exp(X_i \beta). \quad (3.25)$$

The following results directly extend to clusters of any size. For simplicity, the subscript i will be suppressed for the remainder of this section.

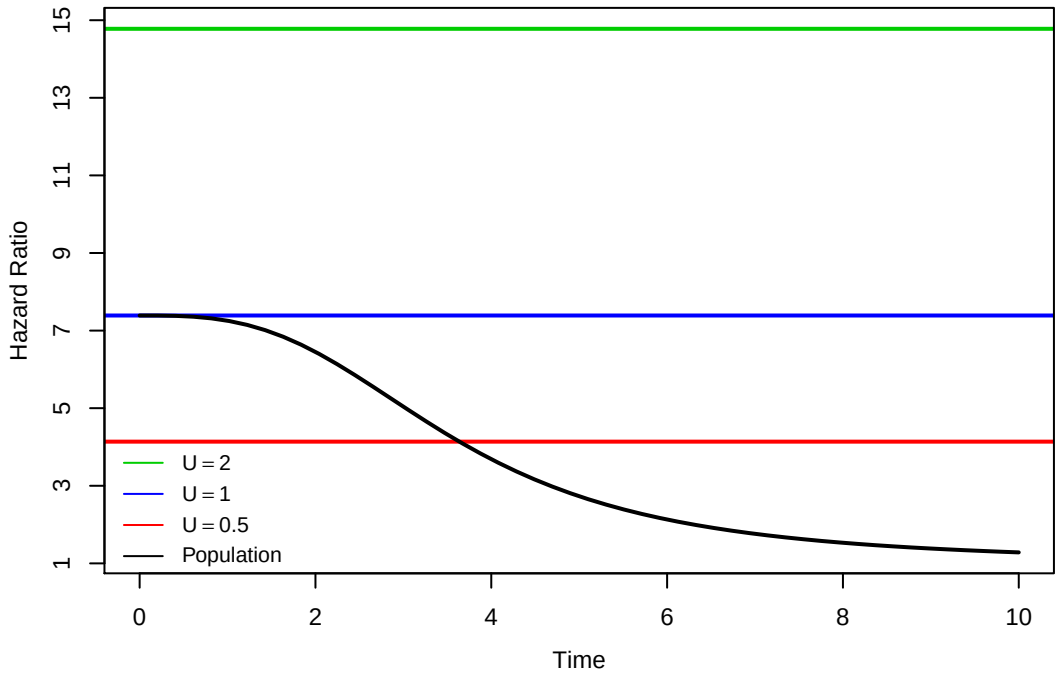

Figure 3.4: Marginal and conditional hazard ratios; where U follows a gamma distribution with mean 1 and variance $\theta = 1$, $\Lambda_0(t) = (t/7)^3$ and $\beta = 2$.

Assuming the frailty model (3.25) is the true model, the marginal or population hazard ratio is given by,

$$\frac{\lambda(t|X=1)}{\lambda(t|X=0)} = \exp(\beta)\frac{\lambda_0(t)E(U|T>t, X=1)}{\lambda_0(t)E(U|T>t, X=0)}. \tag{3.26}$$

It is clear from the above formula that the population hazard ratio will be time dependent except under specific circumstances. Whereas the conditional hazard ratio,

$$\frac{\lambda(t|U=u, X=1)}{\lambda(t|U=u, X=0)} = \frac{\lambda_0(t)u\exp(\beta)}{\lambda_0(t)u} = \exp(\beta). \tag{3.27}$$

does not depend on time. Note from (3.27) that the hazard ratios for frailty models have a cluster-specific interpretation where the hazard ratio refers to comparisons within the same cluster where observations share the same frailty.

To see the difference between the marginal and conditional hazard ratios, assume a gamma distribution with mean 1 and variance θ for U (3.4). Then using (3.16) the hazard ratio for the population is,

$$\exp(\beta)\frac{1+\theta\Lambda_0(t)}{1+\theta\Lambda_0(t)\exp(\beta)}. \tag{3.28}$$

With a gamma frailty distribution, the population hazard ratio is generally time dependent and is only independent of time if $\theta = 0$ or $\beta = 0$. As t increases the hazard ratio decreases to one. Like with expected frailty in section 3.1.2 the decrease in the hazard ratio is increased for large θ, large $\Lambda_0(t)$ or large β. On the other hand if these factors are small then the decrease is modest. Figure 3.4 graphs the marginal and conditional hazard ratios, assuming a Weibull distribution for T with $\lambda_0(t) = (t/7)^3$ and $\beta = 2$. It is clear from this figure that mistakes can be made about the effects of treatment if the frailty term is ignored.

For other frailty distributions there can be even greater differences between the marginal and conditional hazard ratios. For example, it is possible to have a crossover effect where the population hazard ratio changes from being greater than one to less than one as time goes on. In other words, the group that is at high risk of having an event becomes the low risk group as time increases. Consider the compound Poisson distribution (Aalen, 1992) with density function,

$$\begin{aligned} f(u) &= \exp\left(-\alpha(1-\gamma)\left(\frac{u}{\mu}-\frac{1}{\gamma}\right)\right) \\ &\times \frac{1}{\pi}\sum_{\kappa=1}^{\infty}\frac{(\alpha(1-\gamma))^{\kappa(1-\gamma)}\mu^{\kappa\gamma}}{\gamma^{\kappa}}\frac{\Gamma(\kappa\gamma+1)}{\kappa!}(-u)^{-\kappa\gamma-1}\sin(\kappa\gamma\pi) \end{aligned} \tag{3.29}$$

where $u \geq 0$ and the parameters $\mu, \alpha, \kappa > 0$ and $\gamma < 0$; $E(U) = \kappa\alpha^{\gamma-1}$ and $\text{Var}(U) = \kappa(1-\gamma)\alpha^{\gamma-2}$. An interesting feature of this distribution is that it allows for a subgroup of observations that never experience the event, a subgroup where the frailty is zero. Making the usual restrictions $E(U) = 1$ and $\text{Var}(U) = \frac{1-\gamma}{\alpha} = \theta$, the population hazard ratio is,

$$\frac{\lambda(t|X=1)}{\lambda(t|X=0)} = e^{\beta}\frac{\left(1+\frac{\theta}{1-\gamma}\Lambda_0(t)\right)^{1-\gamma}}{\left(1+\frac{\theta}{1-\gamma}\Lambda_0(t)e^{\beta}\right)^{1-\gamma}}. \tag{3.30}$$

When $t = 0$ the hazard ratio is e^{β} and converges towards $e^{\gamma\beta}$ as $t \to \infty$. Under this distribution it is possible to have a crossover effect of the hazard ratios. For example, assuming a Weibull distribution for T with $\Lambda_0(t) = (t/7)^3$, $\beta = 2$, variance $\theta = 1$ and $\gamma = -0.5$, the hazard ratio at $t = 0$ is $\exp(\beta) = 7.4$ while when $t = 10$ it is 0.62 and ultimately converges to 0.37 when $t = \infty$. Thus the high risk group at the beginning of the study becomes the low risk group by the end of the study.

3.2 UNIVARIATE FRAILTY MODEL

If there is only one observation for each cluster, $n_i = 1$ for all i, then (3.1) is the hazard function for a univariate frailty model,

$$\lambda_i(t|u_i) = u_i\lambda_0(t)\exp(X_i^T\beta) \tag{3.31}$$

where the subscript j is no longer necessary. This model is used to account for unobserved heterogeneity that can be the result of not including important covariates in the analysis. With a homogeneous population all individuals have the same risk of experiencing the event, thus it is reasonable to assume a common hazard function for everyone. However, if important covariates are not included in the analysis then the population is heterogeneous and individuals have different risks of experiencing the event (Wienke, 2011). In this case, a different hazard function needs to be fit for each individual; including the frailty effect U_i in (3.31) changes the hazard function for each individual and accounts for the heterogeneity.

In a heterogeneous study population there are individuals with a high risk of failure and there are those individuals with a low risk of having an event. As time goes by, the individuals that remain in the study tend to be of lower risk. Thus estimates of the individual hazard rate without considering the frailty term will increasingly underestimate the hazard rate as time goes by (section 3.1.3). Therefore, to have unbiased estimates it is important to include a frailty term to account for unobserved heterogeneity.

The EM algorithm (Nielsen et al., 1992, Klein, 1992, Vaida and Xu, 2000), the penalized partial likelihood method (Therneau et al., 2003) and more recently hierarchical likelihood

(Ha et al., 2010) can all be used to estimate the univariate frailty model. Of these three approaches, simulation studies suggest that under certain situations the method using hierarchical likelihood will give the least biased estimates (Ha et al., 2010). To guarantee identifiability of the parameters when fitting the model it is necessary to have a sufficiently variable predictor; for the shared frailty model (section 3.3) within-cluster associations provide sufficient information to identify the parameters (Barker and Henderson, 2005). This work does not focus on univariate frailty models, for a more complete discussion of this model see Wienke (2011).

3.3 SHARED FRAILTY MODEL

When the cluster sizes are greater than 1, $n_i > 1$, (3.1) is the hazard function for the shared frailty model,

$$\lambda_{ij}(t|u_i) = u_i \lambda_0(t) \exp(X_{ij}^T \beta). \tag{3.32}$$

The shared frailty model is used to model dependent event times. Shared frailty models are based on two main assumptions (Huang and Wolfe, 2002). First, the frailty effect U_i is shared by each member of the cluster. Second, within each cluster censoring times and event times are independent.

There are several approaches for fitting the shared frailty model. Since the frailties are considered unobserved covariates, Nielsen et al. (1992) and Klein (1992) proposed using the EM algorithm assuming a gamma distribution for U_i to maximizes the observed data likelihood or marginal likelihood where the frailty is integrated out,

$$L_m(\beta, \theta, \lambda_0) = \int L(\beta, \lambda_0|u) f(u|\theta)\, du \tag{3.33}$$

where $f(u|\theta)$ is the frailty density function and

$$L(\beta, \lambda_0|u) = \prod_{ij} \left(u_i \lambda_0(t_{ij}) \exp(X_{ij}^T \beta)\right)^{\delta_{ij}} \exp\left(-u_i \Lambda_0(t_{ij}) \exp(X_{ij}^T \beta)\right) \tag{3.34}$$

is the likelihood function conditional on the frailties $u = (u_1, u_2, \ldots, u_n)$, where t_{ij} is the observed event time and δ_{ij} is the event indicator (1 if an event occurs and 0 otherwise) for the jth observation in cluster i.

McGilchrist and Aisbett (1991) and McGilchrist (1993) developed a procedure using the partial likelihood and assuming a lognormal frailty distribution. More recently, Vaida and Xu (2000), Garnst et al. (2009) and Ripatti et al. (2002) used the EM algorithm to fit models with multivariate lognormal frailties. Ripatti and Palmgren (2000) and Therneau et al. (2003) present a penalized partial likelihood (PPL) approach; for gamma frailties the PPL estimator is the exact same as the EM algorithm and for lognormal frailties the PPL estimator is equivalent to the maximum likelihood estimator presented by McGilchrist (1993). Liu and Huang (2008) has suggested using Gaussian quadrature to fit shared as well as joint frailty models assuming a piecewise constant baseline hazard function.

Ha et al. (2001) and Ha and Lee (2003) extended the h-likelihood (Lee and Nelder, 1996) techniques to estimating regression coefficients under frailty models with gamma and lognormal frailties for parametric and non-parametric baseline hazard functions. The h-likelihood method provides a unified procedure for fitting models that contain both fixed parameters and unobserved random effects. Often, the random effects are integrated out and the marginal likelihood is used for inference using the EM algorithm. A disadvantage of the EM algorithm is that the procedure can be computationally intensive if there is no closed form expression of the marginal survival function, such as for the lognormal frailty distribution. In this case, Markov chain Monte Carlo methods are used to numerically integrate the conditional expectation of the frailty given the observed data. Moreover the estimated standard errors of the regression coefficients are not readily available. Instead they need to be computed using Louis' formula (Louis, 1982). The advantage of the h-likelihood is that it does not require any integration because the h-likelihood includes the random effects in the estimation procedure and does not use the marginal distribution. Since the frailties are not integrated out, this approach also allows for direct inference on the random effects. It was shown in Ha et al. (2001) that for a gamma frailty distribution, the h-likelihood estimates of β and v conditional on θ are the same as the estimates returned by the EM algorithm.

The penalized partial likelihood (PPL) and h-likelihood both give the same estimates of β and v for fixed θ. The difference between the two methods is in the estimation of θ. For lognormal frailty models PPL uses an approximate marginal maximum likelihood estimator for θ. Whereas h-likelihood uses an approximate restricted likelihood estimator for θ. Furthermore the h-likelihood uses a higher order approximation which can lead to less biased estimates of the frailty parameters θ (Ha and Lee, 2003).

3.4 CAUSE-SPECIFIC HAZARD FRAILTY MODEL

The shared frailty model assumes that within clusters censoring times are independent of event times. When competing risks are present this assumption is no longer reasonable, since subjects who experience a competing risks event are censored informatively. To avoid biased results the information from informatively censored subjects needs to be included.

The cause-specific hazard frailty model is a generalization of the shared frailty model that allows for competing risks as well as independent censoring. A similar model was introduced by Huang and Wolfe (2002). Suppose there are $k = 1, 2, \ldots, m$ event types and assume V_i is a random variable from a univariate distribution with parameter θ. Then the cause-specific hazard function conditional on the frailty for the jth observation in cluster i who failed from cause k is,

$$\lambda_{ijk}(t|v_i) = \lambda_{0k}(t) \exp(X_{ij}^T \beta_k + v_i) \tag{3.35}$$

where $\lambda_{0k}(t)$ is the baseline hazard function for event type k and $\beta_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kp})^T$ is a $p \times 1$ vector of fixed parameters for event k. Let $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ be a $mp \times 1$ vector of all the regression coefficients for all event types. Similarly let $\lambda_0 = (\lambda_{01}, \lambda_{02}, \ldots, \lambda_{0m})$ denote the collection of all baseline hazard functions. If there is only one event type $m = 1$ then (3.35) is simply the shared frailty model (3.32).

Model (3.35) simply accounts for correlated event times but there are some limitations to this model which may cause it to be a poor fit in some situations. First, the model assumes that the effect of the frailty is the same for every type of event within a cluster. However, this

assumption may not be reasonable. There can easily be instances where on average, subjects who experience one type of event are more frail and subjects who experience a second event type are less frail.

Another limitation is that (3.35) only allows for positive association within clusters. If the true value of the frailty is less than one then everyone in the cluster experiences an event of any type at an earlier time compared to when the frailty is greater than one where everyone in the cluster experiences any event type at a later time. Thus there is a positive association between observations within a cluster. However there may be cases where there is actually a negative association within a cluster. For example, reducing the risk of dying from cancer can increase the risk of dying from some other disease.

To correct for these limitations, a variation of the bivariate frailty model (Xue and Brookmeyer, 1996) is used in which there is a random effect for each event type,

$$\lambda_{ijk}(t|v_i) = \lambda_{0k}(t)\exp(X_{ij}^T\beta_k + v_{ik}), \tag{3.36}$$

where v_{ik} is the random effect for event k in cluster i. With this model each cluster will have m random effects, one for each event type.

A multivariate distribution needs to be assumed for the random effects, $(V_{i1}, V_{i2}, \ldots, V_{im})$. A natural choice is the multivariate normal distribution with mean 0 and $m \times m$ covariance matrix Σ. Different covariance patterns can be assumed for Σ, such as for $m = 3$,

1. Independent, no correlation between event types

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix}$$

2. Exchangeable, correlation is the same between every event type

$$\Sigma = \begin{pmatrix} \sigma_{11} & \rho\sigma_{11}\sigma_{22} & \rho\sigma_{11}\sigma_{33} \\ \rho\sigma_{11}\sigma_{22} & \sigma_{22} & \rho\sigma_{22}\sigma_{33} \\ \rho\sigma_{11}\sigma_{33} & \rho\sigma_{22}\sigma_{33} & \sigma_{33} \end{pmatrix}$$

3. Unstructured, separate correlation between every event type

$$\Sigma = \begin{pmatrix} \sigma_{11} & \rho_{12}\sigma_{11}\sigma_{22} & \rho_{13}\sigma_{11}\sigma_{33} \\ \rho_{12}\sigma_{11}\sigma_{22} & \sigma_{22} & \rho_{23}\sigma_{22}\sigma_{33} \\ \rho_{13}\sigma_{11}\sigma_{33} & \rho_{23}\sigma_{22}\sigma_{33} & \sigma_{33} \end{pmatrix}$$

It is also possible to assume a constant variance $\sigma_{11} = \sigma_{22} = \sigma_{33}$ for all event types rather than the more general structures given above.

An advantage of a multivariate normal frailty distribution is that it makes it easy to estimate the correlation between the random effects for different event types by estimating the variance components of Σ. This also provides some insight into the association between different types of events. If there is a strong association between two random effects then it is reasonable to assume that there is also a strong association between the two corresponding event types. A positive (negative) correlation indicates that observations with a large random effect for one event type will also have a large (small) random effect for a different event type. Thus their risk for a different event will be high (low), because large (small) random effects increase (decrease) the risk of failure for a cluster.

Since there is an association between the random effects it is not possible to model each event type separately as is normally done when modeling the cause-specific hazard rates without clustering. Instead the effects for both event types as well as all of the random effects need to be estimated jointly. Of course if the random effects are independent then there are no competing risks and the shared frailty model can be fit for each event type.

For the jth observation in the ith cluster, let $T_{ijk}, k = 1, \ldots, m$ denote the time to event for event type k and let C_{ij} denote the independent censoring time. Then the observed event time is $T_{ij} = \min(T_{ij1}, T_{ij2}, \ldots, T_{ijm}, C_{ij})$ and define the event indicator $\delta_{ijk} = 1$ if $T_{ij} = T_{ijk}$ and 0 otherwise. Then the conditional likelihood for cluster i given the frailty v_{ik} is

$$L_i(\beta, \lambda_0 | v_{ik}) = \prod_{k=1}^{m} \prod_{j=1}^{n_i} \left[\lambda_{0k}(t_{ij}) e^{X_{ij}^T \beta_k + v_{ik}}\right]^{\delta_{ijk}} \exp\left(-\Lambda_{0k}(t_{ij}) e^{X_{ij}^T \beta_k + v_{ik}}\right). \tag{3.37}$$

The conditional likelihood for model (3.35) is very similiar. Notice that the above notation can be used for clustered individuals as well as recurrent event times and is general enough to have recurrent event times with different event types.

Like all frailties models the effects of the coefficients β have a cluster-specific interpretation. In the presence of competing risks the effects of the cause-specific hazard frailty model represent the pure effect when the other types of competing events are removed and no longer exist. This approach can be useful for understanding the biological process for a specific outcome. However in a clinical setting this interpretation is not very helpful. In this case it would be more useful to know the effect of the covariate in the presence of other competing risks. This kind of interpretation is given by the subhazard frailty model.

3.5 SUBHAZARD FRAILTY MODEL

An alternative approach for modeling competing risks data is to look at the effect of the covariates on the cumulative incidence function by modeling the subhazard function (Fine and Gray, 1999). This approach estimates the probability of an event of interest for various values of covariates when there are competing risks. Depending on the purpose of the analysis this interpretation can be more appropriate than using the cause-specific hazard frailty, which estimates the covariate effect assuming the other event types are inoperative.

Frailty terms were first added to subhazard models by Katsahian et al. (2006) to account for clustering in multi-center clinical trials. The subhazard frailty model is defined using the hazard function of the subdistribution,

$$\gamma_{ijk}(t|v_{ik}) = \gamma_{0k}(t)\exp(X_{ij}^T\beta_k + v_{ik}) \tag{3.38}$$

where γ_{0k} is the baseline subhazard function. It is possible to assume a multivariate distribution for the frailties like in the previous section and estimate the covariates effects jointly for all event types. However, this work will follow Katsahian et al. (2006) and only model the subhazard for the event of interest, when $k = 1$. This model implicitly assumes that the frailty effect for the event of interest is independent of the frailty effects for the other types of events. Since the subhazard frailty model only considers one event type the subscript k will not be used. Also following Katsahian et al. (2006) assume the random effect V_i is a normal random variable with mean 0 and variance θ.

Let T_{ij} denote the observed event time for the jth observation in the ith cluster and let δ_{ij} define the corresponding event indicator where $\delta_{ij} = 1$ if the event of interest occurred and 0 otherwise. Also let $Z_{ij} = (Z_{ij1}, Z_{ij2}, \ldots, Z_{ijn})^T$ denote a cluster indicator vector where $Z_{ijq} = 1$ if $i = q$ and 0 otherwise. For a vector of frailties, $v = (v_1, v_2, \ldots, v_n)$ it follows that $v_i = Z_{ij}^T v$. Then the form of the conditional partial likelihood for the event of interest $k = 1$ given v is,

$$L_1(\beta|v) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left(\frac{\exp(X_{ij}^T \beta + Z_{ij}^T v)}{\sum_{r \in \mathcal{R}_{ij}} w_{ij}(t_r) \exp(X_r^T \beta + Z_r^T v)} \right)^{\delta_{ij}}. \tag{3.39}$$

This partial likelihood resembles the partial likelihood used for the proportional hazards model with two notable exceptions. First, the risk set $\mathcal{R}_{ij}$ is all individuals who have not experienced an event by time t_{ij} and those who experienced a competing event by time t_{ij},

$$\mathcal{R}_{ij} = \{r : (T_{ij} \leq T_r) \cup (T_{ij} > T_r \cap \delta_{ij} = 0)\}. \tag{3.40}$$

Note that both ij and r denote a single observation from the total sample size N.

Second, there are weights $w_{ij}(t_r)$ are defined as,

$$w_{ij}(t_r) = \frac{\hat{G}(t_r)}{\hat{G}(\min(t_r, t_{ij}))} \tag{3.41}$$

where $\hat{G}$ is the Kaplan-Meier estimator of the survival function for the censoring distribution; the estimator is calculated treating censored observations as event times and event times as censored observations, i.e. by using the data $\{T_{ij}, 1 - \delta_{ij}\}$. If there is no censoring then $w_{ij}(t_r) = 1$ for all r.

When censoring is present, the weights equal one if individuals have not had an event. For individuals who failed from a competing event, the weights decrease over time. Therefore, individuals who have not failed fully contribute to the partial likelihood, while those who failed from a competing cause only partially contribute to the partial likelihood. Their contribution is weighted according to their probability of being censored. Ideally, they should only remain in the risk set until they are censored, but the independent censoring time is precluded by the occurrence of the competing event. When a subject experiences the event of

interest or is right censored they are no longer in the risk set and do not have any additional contribution to the partial likelihood.

Katsahian et al. (2006) first proposed an estimation procedure adapted from McGilchrist (1993) that uses the restricted maximum likelihood approach. Later Katsahian and Boudreau (2011) proposed another estimation procedure that uses the PPL approach. An advantage of this recent method is that it can be done using existing software. H-likelihood and PPL both give the same estimates of β and v for fixed θ (Ha and Lee, 2003). Since h-likelihood uses a higher order approximation to estimate θ it is possible that h-likelihood will give more accurate estimates of the frailty parameter θ for the subhazard frailty model compared to the PPL method proposed by Katsahian and Boudreau (2011). This will be investigated by extensive simulations in Chapter 5.

4.0 HIERARCHICAL LIKELIHOOD

Lee and Nelder (1996) first proposed using the hierarchical likelihood (h-likelihood) to fit hierarchical generalized linear models. The h-likelihood method provides a unified procedure for fitting models that contain both fixed parameters and unobserved random effects. The biggest advantage of h-likelihood is that it does not require any integration, which can often be a computational hurdle depending on the model and estimation method.

This chapter derives the h-likelihood estimators for the cause-specific hazard frailty model and the subhazard frailty model assuming either a univariate or multivariate normal distribution for the random effects. First sections 4.1 and 4.2 derive the estimation procedure for the cause-specific hazard frailty model. Next section 4.3 discusses using h-likelihood to fit the subhazard frailty model. Finally, section 4.4 discusses techniques for inference and model selection.

Given the similarities between fitting a shared frailty model, cause-specific hazard frailty model and subhazard frailty model it is possible to write a general program that will fit each model. The appendix contains an `R` program that performs the estimation and inference discussed in this chapter (R Development Core Team, 2010).

4.1 ESTIMATING THE CAUSE-SPECIFIC HAZARD MULTIVARIATE FRAILTY MODEL

The purpose of this section is to derive the h-likelihood estimators of the fixed effects β and the frailty parameter θ as well as predict the random effects v for the semiparametric cause-specific hazard frailty model (3.36), where the functional form of the baseline hazard is not

specified and the random effects are multivariate normal. The following estimation procedure is similar to Ha et al. (2011) who analyzed multi-center clinical trial data accounting for the between-center variation and the treatment effect across centers by assuming a bivariate normal frailty distribution. For $m = 1$ this estimation method reduces to the shared frailty case (Ha et al., 2001, Ha and Lee, 2003). To simplify the notation, assume there are just two event types $k = 1, 2$, the results easily generalize to m event types.

First, define the h-likelihood function. Suppose there are $i = 1, 2, \ldots, n$ clusters where each cluster has $j = 1, 2, \ldots, n_i$ observations, so that the total sample size is $N = \sum_{i=1}^{n} n_i$. The two indices i and j denote a unique observation from the overall sample of size N. Following Ha et al. (2001), the contribution of the jth observation in the ith cluster to the h-likelihood for event k is given by the log of the joint density function of $(T_{ij}, \delta_{ijk}, v_{ik})$ written as a function of the parameters for event k, $(\beta_k, \lambda_{0k}, \theta)$,

$$h_{ijk} = h_{ijk}(\beta_k, \lambda_{0k}, \theta; t_{ij}, \delta_{ijk}, v_{ik}) \quad = \quad \log\left[f_{ijk}(t_{ij}, \delta_{ijk}; \beta_k, \lambda_{0k}|v_{ik}) f_i(v_i; \theta)\right] \tag{4.1}$$

where f_{ijk} is the conditional density function of T_{ij} and δ_{ijk} given $V_{ik} = v_{ik}$ with parameters (β_k, λ_{0k}) and f_i is the density function of $V_i = (V_{i1}, V_{i2})$ with parameter θ. The conditional density function of (T_{ij}, δ_{ijk}) given $V_{ik} = v_{ik}$ is

$$f_{ijk}(t_{ij}, \delta_{ijk}; \beta_k, \lambda_{0k}|v_{ik}) = \left[\lambda_{0k}(t_{ij}) e^{X_{ij}^T \beta_k + v_{ik}}\right]^{\delta_{ijk}} \exp\left(-\Lambda_{0k}(t_{ij}) e^{X_{ij}^T \beta_k + v_{ik}}\right). \tag{4.2}$$

Assume $V_i = (V_{i1}, V_{i2})$ follows a bivariate normal distribution with mean 0 and covariance matrix Σ. Then the density function is,

$$f_i(v_i) = \frac{1}{(2\pi)|\Sigma|^{1/2}} \exp\left(-\frac{1}{2} v_i^T \Sigma^{-1} v_i\right). \tag{4.3}$$

Let $\theta = (\sigma_{11}, \sigma_{22}, \sigma_{12})^T$ be the parameter vector for $f_i(v_i)$ that contains the variance components of Σ. Notation note, it will be helpful to distinguish between random effects associated with a cluster and random effects for a particular event type. So, let $V_i = (V_{i1}, V_{i2})^T$ denotes the random effects vector for both event types for cluster i and let $V_k = (V_{1k}, V_{2k}, \ldots, V_{nk})^T$ denote a n-dimension vector of the random effects from all clusters just for event k; the random effect V_{ik} is the effect for event k in cluster i. Also, let $V = (V_{11}, V_{21}, \ldots, V_{n1}, V_{12}, V_{22}, \ldots, V_{n2})^T$ be a $2n$-dimensional vector of all random effects,

for all clusters and event types. Notice that the random effects are arranged by event type so that all of the random effects for the same event type are adjacent. The structure of V is important later on in this chapter only because it determines how the observed information matrices are structured. Other forms of V are allowable, but will require rearranging terms. Similarly, let $\beta = (\beta_1, \beta_2)^T$ be a vector of regression coefficients for both event types and let $\lambda_0 = (\lambda_{01}, \lambda_{02})$ be a collection of all the baseline hazards.

Since event times within a cluster are conditionally independent given the frailty $V_i = v_i$ and the frailties V_i are independent and identically distributed random variables, the h-likelihood for the cause-specific hazard frailty model is,

$$h(\beta, \lambda_0, v, \theta) = \sum_{ijk} h_{ijk} = \sum_{ijk} l_{ijk}(\beta_k, \lambda_{0k}; t_{ij}, \delta_{ijk}|v_{ik}) + \sum_i l_i(\theta; v_i) \tag{4.4}$$

where

$$l_{ijk}(\beta_k, \lambda_{0k}; t_{ij}, \delta_{ijk}|v_{ik}) = \delta_{ijk}\left(\log \lambda_{0k}(t_{ij}) + x_{ij}^T\beta_k + v_{ik}\right) - \Lambda_{0k}(t_{ij}) \exp\left(x_{ij}^T\beta_k + v_{ik}\right)$$

is the log of the conditional density function for T_{ij} and δ_{ijk} given $V_{ik} = v_{ik}$ and

$$l_i(\theta; v_i) = -\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}v_i^T\Sigma^{-1}v_i$$

is the log of the bivariate normal density function for V_i with parameter θ.

No parametric form is assumed for the baseline hazard function $\lambda_{0k}(t)$. Instead assume that the cumulative baseline hazard function for event type k is a step function with jumps at the observed event times,

$$\Lambda_{0k}(t) = \sum_{r\,:\,t_{(kr)} \le t} \lambda_{0kr} \tag{4.5}$$

where $t_{(k1)} < t_{(k2)} < \ldots < t_{(kD_k)}$ denote the D_k ordered unique event times for type k events among all of the t_{ij}'s that refer to the time of a type k event and $\lambda_{0kr} = \lambda_{0k}(t_{(kr)})$. Also let d_{kr} be the number of events that occur at time $t_{(kr)}$.

As the number of type k events increases, the number of nuisance parameters, λ_{0kr}, $r = 1, 2, \ldots, D_k$, also increases. This requires the estimation of a high dimensional baseline hazard function λ_{0k}. Therefore, following Ha et al. (2001) the profile h-likelihood is used

where λ_{0k} and Λ_{0k} are removed. Murphy and van der Vaart (2000) give a general justification for using a semi-parametric profile likelihood for statistical inference.

To calculate the profile h-likelihood, first rewrite the h-likelihood (4.4) so that it will be easier to differentiate with respect to λ_{0kr} for fixed β_k, v_k and θ. Let Z_{ij}=$(Z_{ij1}, Z_{ij2}, \ldots, Z_{ijn})^T$ be a $n \times 1$ cluster indicator vector where $Z_{ijq} = 1$ if $i = q$ and 0 otherwise; let Z be a $N \times n$ matrix whose ij row is Z_{ij}^T. Then $v_{ik} = Z_{ij}^T v_k$ for any j. Replace v_{ik} with $Z_{ij}^T v_k$ in (4.4) and sum over the unique event times for each event type. Then by using (4.5) the h-likelihood (4.4) can be rewritten as,

$$h = \sum_{k=1}^{2}\left[\sum_{r=1}^{D_k} d_{kr}\log\lambda_{0kr} + S_{Xkr}^T\beta_k + S_{Zkr}^T v_k - \lambda_{0kr}\sum_{ij\in\mathcal{R}_{kr}}\exp\left(X_{ij}^T\beta_k + Z_{ij}^T v_k\right)\right] + \sum_{i=1}^{n} l_i(\theta; v_i) \quad (4.6)$$

where, $S_{Xkr}^T = \sum_{ij\in\mathcal{D}_{kr}} X_{ij}^T$ and $S_{Zkr}^T = \sum_{ij\in\mathcal{D}_{kr}} Z_{ij}^T$ are the sums of the vectors X_{ij}^T and Z_{ij}^T over the set $\mathcal{D}_{kr} = \{ij : \delta_{ijk} = 1 \text{ and } t_{ij} = t_{(kr)}\}$ of all individuals who have a type k event at time $t_{(kr)}$ and $\mathcal{R}_{kr} = \{ij : t_{ij} \geq t_{(kr)}\}$ is the risk set at time $t_{(kr)}$, this is the set of all individuals who are still at risk to experience an event.

Using an approach similar to Johansen (1983), by fixing β_k, v_k and θ and maximizing (4.6) as a function of λ_{0kr} gives the nonparametric maximum hierarchical likelihood estimator of λ_{0kr},

$$\hat{\lambda}_{0kr} = \frac{d_{kr}}{\sum_{ij\in\mathcal{R}_{kr}}\exp(X_{ij}^T\beta_k + Z_{ij}^T v_k)}. \quad (4.7)$$

Thus, $\hat{\Lambda}_{0k}(t) = \sum_{r:\, y_{kr}\leq t}\hat{\lambda}_{0kr}$ is an extension of the Breslow (1974) estimator of the baseline cumulative hazard function. Replacing λ_{0kr} with $\hat{\lambda}_{0kr}$ in (4.6) gives the profile h-likelihood as a function of β, v, and θ only,

$$h_p(\beta, v, \theta) = \sum_{kr} S_{Xkr}^T\beta_k + S_{Zkr}^T v_k - d_{kr}\log\left(\sum_{ij\in\mathcal{R}_{kr}}\exp\left(X_{ij}^T\beta_k + Z_{ij}^T v_k\right)\right) - d_{kr} + \sum_{i=1}^{n} l_i(\theta; v_i). \quad (4.8)$$

The maximum hierarchical likelihood estimators (MHLE) of β and v are found by maximizing the profile h-likelihood for fixed θ using the Newton-Raphson method (Tanner, 1996); an iterative procedure that uses the gradient vector and observed information matrix to approximate the points that maximize a likelihood function. In this case, starting at initial

values $\hat{\beta}^{(0)}$ and $\hat{v}^{(0)}$ the approximate maximums are updated iteratively until convergence is achieved by,

$$\begin{pmatrix} \hat{\beta}^{(i+1)} \\ \hat{v}^{(i+1)} \end{pmatrix} = \begin{pmatrix} \hat{\beta}^{(i)} \\ \hat{v}^{(i)} \end{pmatrix} + \left[\begin{pmatrix} \frac{-\partial^2 h_p}{\partial \beta^2} & \frac{-\partial^2 h_p}{\partial \beta \partial v} \\ \frac{-\partial^2 h_p}{\partial v \partial \beta} & \frac{-\partial^2 h_p}{\partial v^2} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial h_p}{\partial \beta} \\ \frac{\partial h_p}{\partial v} \end{pmatrix} \right] \Bigg|_{(\beta, v) = (\hat{\beta}^{(i)}, \hat{v}^{(i)})} \tag{4.9}$$

where $\hat{\beta}^{(i)}$ and $\hat{v}^{(i)}$ represent the estimates of β and v at the ith iteration. Reasonable starting values for $\hat{\beta}^{(0)} = (\hat{\beta}_1^{(0)}, \hat{\beta}_2^{(0)})^T$ are the estimates from the cause-specific hazard model for each event type with no random effects. A possible starting value for $\hat{v}^{(0)}$ is a random sample of size n from the bivariate normal distribution (4.3) with $\theta = \hat{\theta}^{(0)}$, where $\hat{\theta}^{(0)}$ are the initial variance components. Unfortunately, there is no good way to select appropriate starting values for θ, so different values will need to be tried.

First the elements of the gradient vector $(\partial h_p/\partial \beta, \partial h_p/\partial v)^T$are calculated. The kth element of $\partial h_p/\partial \beta = (\partial h_p/\partial \beta_1,\ \partial h_p/\partial \beta_2)^T$ is the derivative of h_p with respect to the regression coefficients for event k,

$$\frac{\partial h_p}{\partial \beta_k} = \sum_{ij} X_{ij}\delta_{ijk} - X_{ij}\hat{\Lambda}_{0k}(t_{ij}) \exp\left(X_{ij}^T\beta_k + Z_{ij}^T v_k\right) \tag{4.10}$$

and $\partial h_p/\partial v = (\partial h_p/\partial v_1,\ \partial h_p/\partial v_2)^T$ is the derivative of h_p with respect to the random effects for each event type where for event k,

$$\frac{\partial h_p}{\partial v_k} = \sum_{ij} Z_{ij}\delta_{ijk} - Z_{ij}\hat{\Lambda}_{0k}(t_{ij}) \exp\left(X_{ij}^T\beta_k + Z_{ij}^T v_k\right) - \sum_{i=1}^{n} v_i \bullet (\sigma_{kk}, \sigma_{12}) \tag{4.11}$$

where $\bullet$ denotes the inner product of two vectors.

Calculating the estimators will be much easier if matrices are used instead of summations. The following matrices and notation are used for the remainder of this section. Let $R_k = (R_1, R_2, \ldots, R_{D_k})$ be a $N \times D_k$ at risk indicator matrix where the ijth element in column r is one if $t_{ij} \geq t_{kr}$ and zero otherwise. Define E_k as a $N \times 1$ type k event indicator vector with ijth element δ_{ijk}. Let M_k be a $N \times N$ diagonal matrix with elements $\hat{\Lambda}_{0k}(t_{ij}) \exp(X_{ij}^T\beta_k + Z_{ij}^T v_k)$, let N_k be a $N \times N$ diagonal matrix with elements $\exp(X_{ij}^T\beta_k + Z_{ij}^T v_k)$ and let C_k be a diagonal $D_k \times D_k$ matrix where the rth element is $\hat{\lambda}_{0kr}/d_{kr}$. Finally, let I_n be a $n \times n$ identity matrix

and let $\otimes$ denote the Kronecker product. Recall that X is a $N \times p$ matrix of p covariates and Z is a $N \times n$ cluster indicator matrix.

Using this notation the gradient vector can be rewritten in a more compact form. First equation (4.10) can be expressed as

$$\frac{\partial h_p}{\partial \beta_k} = X^T(E_k - M_k). \tag{4.12}$$

Second the derivative of h_p with respect to all random effects v is,

$$\frac{\partial h_p}{\partial v} = \begin{pmatrix} Z^T(E_1 - M_1) \\ Z^T(E_2 - M_2) \end{pmatrix} - (\Sigma^{-1} \otimes I_n). \tag{4.13}$$

Next the observed information matrix H of β and v for fixed θ is calculated. First, define $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ as block diagonal matrices such that,

$$\mathbf{X} = \begin{pmatrix} X & \mathbf{0} \\ \mathbf{0} & X \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} Z & \mathbf{0} \\ \mathbf{0} & Z \end{pmatrix} \quad \text{and} \quad \mathbf{W} = \begin{pmatrix} W_1 & \mathbf{0} \\ \mathbf{0} & W_2 \end{pmatrix} \tag{4.14}$$

where $\mathbf{0}$ is a conformable matrix of zeros and $W_k = W_k(\beta_k, v_k) = M_k - N_k R_k C_k (R_k N_k)^T$ for $k = 1, 2$. Then the observed information matrix H is a large $(mp + mn) \times (mp + mn)$ matrix defined as,

$$H = H(\beta, v, \theta) = \begin{pmatrix} \mathbf{X}^T\mathbf{W}\mathbf{X} & \mathbf{X}^T\mathbf{W}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{W}\mathbf{X} & \mathbf{Z}^T\mathbf{W}\mathbf{Z} + Q \end{pmatrix}. \tag{4.15}$$

where the elements $\mathbf{X}^T\mathbf{W}\mathbf{X}$, $\mathbf{X}^T\mathbf{W}\mathbf{Z}$, $\mathbf{Z}^T\mathbf{W}\mathbf{X}$ and $\mathbf{Z}^T\mathbf{W}\mathbf{Z}$ are also block diagonal matrices that correspond to partitions of H. That is,

$$\begin{aligned} \frac{-\partial^2 h_p}{\partial \beta^2} &= \mathbf{X}^T\mathbf{W}\mathbf{X} \\ \frac{-\partial^2 h_p}{\partial \beta \partial v} &= \mathbf{X}^T\mathbf{W}\mathbf{Z} \\ \frac{-\partial^2 h_p}{\partial v \partial \beta} &= \mathbf{Z}^T\mathbf{W}\mathbf{X} \\ \frac{-\partial^2 h_p}{\partial v^2} &= \mathbf{Z}^T\mathbf{W}\mathbf{Z} + Q \end{aligned}$$

where Q is a $n \times n$ matrix that is the negative second derivative of the log of the joint density function for all random effects with respect to the vector v,

$$Q(v, \theta) = -\frac{\partial^2}{\partial v^2} \sum_{i=1}^{n} l_i(\theta; v_i) = \Sigma^{-1} \otimes I_n. \tag{4.16}$$

The next step is to find the MHLE of θ by maximizing Lee and Nelder (1996) adjusted profile h-likelihood (Ha et al., 2001, Ha and Lee, 2003),

$$h_A(\theta) = \left[h_p - \frac{1}{2} \log(\det(H/2\pi)) \right] \Big|_{(\beta, v) = (\hat{\beta}(\theta), \hat{v}(\theta))} \tag{4.17}$$

given $\hat{\beta} = \hat{\beta}(\theta)$ and $\hat{v} = \hat{v}(\theta)$, the current estimates of β and v conditional on θ. The adjusted profile h-likelihood is used to approximate the restricted likelihood of θ that takes into account the estimation of β and v. The Newton-Raphson method is also used to find $\hat{\theta}$ the MHLE of θ. This requires finding the first and second derivatives of h_A with respect to every variance component of $\theta = (\sigma_{11}, \sigma_{22}, \sigma_{12})^T$; let θ_q and θ_s denote the qth and sth components of θ respectively, for $q, s = 1, 2, 3$. Two identities from matrix calculus will be useful here (Searle et al. (1992) appendix M). For a matrix A and scalar x,

$$\begin{aligned} \frac{\partial}{\partial x} \log |A| &= \text{trace} \left(A^{-1} \frac{\partial A}{\partial x} \right) \\ \frac{\partial A^{-1}}{\partial x} &= -A^{-1} \frac{\partial A}{\partial x} A^{-1}. \end{aligned}$$

Using the properties of determinants (4.17) can be rewritten as

$$h_A = \hat{h}_p - \frac{1}{2} \log(\det(\hat{H})) + \frac{(p+n)}{2} \log(2\pi) \tag{4.18}$$

where $\hat{h}_p = h_p(\hat{\beta}(\theta), \hat{v}(\theta), \theta)$ and $\hat{H} = H(\hat{\beta}(\theta), \hat{v}(\theta), \theta)$ are the profile h-likelihood and observed information matrix evaluated at the current estimates of β and v, respectively. Then using the above results from matrix calculus, the qth component of the gradient vector $\partial h_A / \partial \theta$ is

$$\frac{\partial h_A}{\partial \theta_q} = \frac{\partial \hat{h}_p}{\partial \theta_q} - \frac{1}{2} \text{trace} \left(\hat{H}^{-1} \frac{\partial \hat{H}}{\partial \theta_q} \right). \tag{4.19}$$

Furthermore, the element in row q and column s of the 3×3 observed information matrix $\partial^2 h_A / \partial \theta^2$ for the frailty parameter θ is

$$-\frac{\partial^2 h_A}{\partial \theta_q \partial \theta_s} = -\frac{\partial^2 \hat{h}_p}{\partial \theta_q \partial \theta_s} + \frac{1}{2}\text{trace}\left(-\hat{H}^{-1}\frac{\partial \hat{H}}{\partial \theta_q}\hat{H}^{-1}\frac{\partial \hat{H}}{\partial \theta_s} + \hat{H}^{-1}\frac{\partial^2 \hat{H}}{\partial \theta_q \partial \theta_s}\right). \tag{4.20}$$

Partial differentiation of a multivariate function with respect to one variable assumes that the remaining variables of the function are held constant. Since $\hat{\beta}(\theta)$ and $\hat{v}(\theta)$ are functions of θ, it is not appropriate to just use partial derivatives in (4.19) and (4.20). Instead the total derivative should be used. Consider the total derivative of h_A with respect to θ_q,

$$\frac{\partial h_A}{\partial \theta_q} = \frac{\partial h_A}{\partial \theta_q} + \left(\frac{\partial h_A}{\partial \beta}\bigg|_{\beta=\hat{\beta}}\right)\frac{\partial \hat{\beta}}{\partial \theta_q} + \left(\frac{\partial h_A}{\partial v}\bigg|_{v=\hat{v}}\right)\frac{\partial \hat{v}}{\partial \theta_q}. \tag{4.21}$$

The total derivative calculates the derivative of h_A with respect to θ_q where the other arguments of h_A, $\hat{\beta}(\theta_q)$ and $\hat{v}(\theta_q)$, are allowed to depend on θ_q; they do not have to remain constant.

Originally, Lee and Nelder (1996) and Ha et al. (2001) ignored $\partial\hat{\beta}/\partial\theta_q$ and $\partial\hat{v}/\partial\theta_q$ when differentiating $\hat{h}_p$ and $\hat{H}$ with respect to θ because the parameters are asymptotically orthogonal (Lee and Nelder, 1996). However this approach does not work in some cases such as data with binary covariates and small cluster sizes. Following Ha and Lee (2003), $\partial\hat{\beta}/\partial\theta$ is ignored because there is an indirect dependency between $\hat{\beta}$ and θ_q whereas $\partial\hat{v}/\partial\theta_q$ is included because there is a direct dependency between $\hat{v}$ and θ_q; this is clear from (4.12) and (4.13).

Let $\tau = (\beta, v)$ and let $\hat{\tau} = \hat{\tau}(\theta) = (\hat{\beta}(\theta), \hat{v}(\theta))$. First calculate the derivatives in (4.19). Since $\partial h_p/\partial v|_{\tau=\hat{\tau}} = 0$ the total derivative of the first term $\partial\hat{h}_p/\partial\theta_q$ is,

$$\begin{aligned}
\frac{\partial \hat{h}_p}{\partial \theta_q} &= \frac{\partial h_p}{\partial \theta_q}\bigg|_{\tau=\hat{\tau}} + \left(\frac{\partial h_p}{\partial v}\bigg|_{\tau=\hat{\tau}}\right)\left(\frac{\partial \hat{v}}{\partial \theta_q}\right) \\
&= \frac{\partial h_p}{\partial \theta_q}\bigg|_{\tau=\hat{\tau}} \\
&= \sum_{i=1}^{n}\frac{\partial l_i(\theta; \hat{v}_i)}{\partial \theta_q} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\text{trace}\left(\Sigma^{-1}\Sigma_q'\right) + \frac{1}{2}\hat{v}_i^T\left(\Sigma^{-1}\Sigma_q'\Sigma^{-1}\right)\hat{v}_i
\end{aligned} \tag{4.22}$$

where $\Sigma'_q = \partial\Sigma/\partial\theta_q$.

The derivative of the second term $\partial\hat{H}/\partial\theta_q$ in (4.19) is more complicated,

$$\frac{\partial\hat{H}}{\partial\theta_q} = \frac{\partial H}{\partial\theta_q}\bigg|_{\tau=\hat{\tau}} + \left(\frac{\partial H}{\partial v}\bigg|_{\tau=\hat{\tau}}\right)\left(\frac{\partial\hat{v}}{\partial\theta_q}\right). \tag{4.23}$$

The term $\partial\hat{v}/\partial\theta_q$ is calculated following Lee et al. (2006). From h_p given θ_q let $\hat{v}(\theta_q)$ be the solution to $g(\theta_q) = \partial h_p/\partial v|_{\tau=\hat{\tau}} = 0$. Then,

$$\frac{\partial^2 g(\theta_q)}{\partial\theta_q} = \frac{\partial h_p}{\partial v\partial\theta_q}\bigg|_{\tau=\hat{\tau}} + \left(\frac{\partial^2 h_p}{\partial v^2}\bigg|_{\tau=\hat{\tau}}\right)\left(\frac{\partial\hat{v}}{\partial\theta_q}\right) = 0 \tag{4.24}$$

Solving for $\partial\hat{v}/\partial\theta_q$ gives a $2n \times 1$ vector,

$$\begin{aligned}\frac{\partial\hat{v}}{\partial\theta_q} &= \left(-\frac{\partial^2 h_p}{\partial v^2}\bigg|_{\tau=\hat{\tau}}\right)^{-1}\left(\frac{\partial^2 h_p}{\partial v\partial\theta_q}\bigg|_{\tau=\hat{\tau}}\right) \\ &= \left(\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} + Q\right)^{-1}\left(\left[\left(-\Sigma^{-1}\Sigma'_q\Sigma^{-1}\right)\otimes I_n\right]\hat{v}\right)\end{aligned} \tag{4.25}$$

where $\hat{\mathbf{W}}$ is $\mathbf{W}$ evaluated at $(\hat{\beta}, \hat{v}, \theta)$; that is, when $W_k = W_k(\hat{\beta}_k, \hat{v}_k, \theta) = \hat{W}_k$. Now since X and Z are constant matrices that have no dependence on θ it follows that the total derivative of $\partial\hat{H}/\partial\theta_q$ is,

$$\frac{\partial\hat{H}}{\partial\theta_q} = \begin{pmatrix} \mathbf{X}^T\hat{\mathbf{W}}'_q\mathbf{X} & \mathbf{X}^T\hat{\mathbf{W}}'_q\mathbf{Z} \\ \mathbf{Z}^T\hat{\mathbf{W}}'_q\mathbf{X} & \mathbf{Z}^T\hat{\mathbf{W}}'_q\mathbf{Z} + Q'_q \end{pmatrix}. \tag{4.26}$$

where $\hat{\mathbf{W}}'_q = \partial\hat{\mathbf{W}}/\partial\theta_q$ and $Q'_q = \partial Q/\partial\theta_q$. Based on the structure of $\mathbf{W}$, $\hat{\mathbf{W}}'_q$ is found by finding $\partial\hat{W}_k/\partial\theta_q$. Since $\hat{W}_k$ does not depend on θ the total derivative of $\hat{W}_k$ is,

$$\hat{W}'_{kq} = \frac{\partial\hat{W}_k}{\partial\theta_q} = \frac{\partial W_k}{\partial\theta_q}\bigg|_{\tau=\hat{\tau}} + \left(\frac{\partial W_k}{\partial v_k}\bigg|_{\tau=\hat{\tau}}\right)\left(\frac{\partial\hat{v}_k}{\partial\theta_q}\right) = \left(\frac{\partial W_k}{\partial v_k}\bigg|_{\tau=\hat{\tau}}\right)\left(\frac{\partial\hat{v}_k}{\partial\theta_q}\right). \tag{4.27}$$

The $\partial W_k/\partial v_k$ is found by differentiating $W_k(\beta_k, v_k) = M_k - N_k R_k C_k (R_k N_k)^T$ with respect to v_k. Given the structure of v defined earlier, $\partial\hat{v}_1/\partial\theta_q$ is the first n elements of the vector $\partial\hat{v}/\partial\theta_q$ and $\partial\hat{v}_2/\partial\theta_q$ are the last n elements. Since Q does not depend on v the total derivative is not needed to find $\partial Q/\partial\theta_q$ so,

$$Q'_q = \frac{\partial Q}{\partial\theta_q} = \left(-\Sigma^{-1}\Sigma'_q\Sigma^{-1}\right)\otimes I_n. \tag{4.28}$$

Using (4.28) there is a slightly simpler expression for $\partial\hat{v}/\partial\theta_q = \left(\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} + Q\right)^{-1} + Q'_q\hat{v}$.

The next step is to calculate the terms in the observed information (4.20). First,

$$\begin{aligned}
-\frac{\partial^2 \hat{h}_p}{\partial\theta_q\partial\theta_s} &= -\frac{\partial^2 h_p}{\partial\theta_q\partial\theta_s}\bigg|_{\tau=\hat{\tau}} - \left(\frac{\partial^2 h_p}{\partial v\partial\theta_s}\bigg|_{\tau=\hat{\tau}}\right)\left(\frac{\partial\hat{v}}{\partial\theta_s}\right) \\
&= \sum_{i=1}^{n} -\frac{\partial^2 l_i(\theta;\hat{v}_i)}{\partial\theta_q\partial\theta_s} - \left(Q'_q\hat{v}\right)\left(\frac{\partial\hat{v}}{\partial\theta_s}\right) \\
&= \sum_{i=1}^{n}\left(\frac{1}{2}\text{trace}\left(\Sigma^{-1}\Sigma'_s\Sigma^{-1}\Sigma'_q\Sigma^{-1}\right) + \frac{1}{2}\hat{v}_i^T\left(\Sigma^{-1}\Sigma'_s\Sigma^{-1}\Sigma'_q\Sigma^{-1}\right)\hat{v}_i\right. \\
&\qquad \left. + \frac{1}{2}\hat{v}_i^T\left(\Sigma^{-1}\Sigma'_q\Sigma^{-1}\Sigma'_s\Sigma^{-1}\right)\hat{v}_i\right) - Q'_q\hat{v}\left(\frac{\partial\hat{v}}{\partial\theta_s}\right). \qquad (4.29)
\end{aligned}$$

The last term needed to calculate (4.20) is,

$$\frac{\partial^2\hat{H}}{\partial\theta_q\partial\theta_s} = \begin{pmatrix} \mathbf{X}^T\hat{\mathbf{W}}''_{qs}\mathbf{X} & \mathbf{X}^T\hat{\mathbf{W}}''_{qs}\mathbf{Z} \\ \mathbf{Z}^T\hat{\mathbf{W}}''_{qs}\mathbf{X} & \mathbf{Z}^T\hat{\mathbf{W}}''_{qs}\mathbf{Z} + Q''_{qs} \end{pmatrix} \qquad (4.30)$$

where,

$$Q''_{qs} = \frac{\partial^2 Q}{\partial\theta_q\partial\theta_s} = \left(\Sigma^{-1}\Sigma'_s\Sigma^{-1}\Sigma'_q\Sigma^{-1} + \Sigma^{-1}\Sigma'_q\Sigma^{-1}\Sigma'_s\Sigma^{-1}\right)\otimes I_n \qquad (4.31)$$

and $\hat{\mathbf{W}}''_{qs} = \partial\hat{\mathbf{W}}'_q/\partial\theta_s$. Like earlier, $\hat{\mathbf{W}}''_{qs}$ is found by finding $\partial^2\hat{W}_k/\partial\theta_q\partial\theta_s$ for $k = 1, 2$,

$$\hat{W}''_{kqs} = \frac{\partial^2\hat{W}_k}{\partial\theta_q\partial\theta_s} = \left[\left(\frac{\partial^2 W_k}{\partial v_k^2}\bigg|_{\tau=\hat{\tau}}\right)\frac{\partial\hat{v}_k}{\partial\theta_q}\right]\frac{\partial\hat{v}_k}{\partial\theta_s} + \left(\frac{\partial W_k}{\partial v_k}\bigg|_{\tau=\hat{\tau}}\right)\frac{\partial^2\hat{v}_k}{\partial\theta_q\partial\theta_s} \qquad (4.32)$$

where,

$$\begin{aligned}
\frac{\partial^2\hat{v}}{\partial\theta_q\partial\theta_s} &= \left(\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} + Q\right)^{-1}\left(\mathbf{Z}^T\hat{\mathbf{W}}'_s\mathbf{Z} + Q'_s\right)\left(\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} + Q\right)^{-1}Q'_q\hat{v} \\
&\qquad - \left(\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} + Q\right)^{-1}\left[Q''_{qs}\hat{v} + Q'_q\frac{\partial\hat{v}}{\partial\theta_s}\right] \\
&= -\left(\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} + Q\right)^{-1}\left[\left(\mathbf{Z}^T\hat{\mathbf{W}}'_s\mathbf{Z} + Q'_s\right)\frac{\partial\hat{v}}{\partial\theta_q} + Q''_{qs}\hat{v} + Q'_q\frac{\partial\hat{v}}{\partial\theta_s}\right] \qquad (4.33)
\end{aligned}$$

is a $2n \times 1$ vector and $\partial^2\hat{v}_k/\partial\theta_q\partial\theta_s$ is the first n elements of (4.33) if $k = 1$ and the second n elements if $k = 2$. The $\partial^2 W_k/\partial v^2$ is found by twice differentiating $W_k(\beta_k, v_k) = M_k - N_k R_k C_k (R_k N_k)^T$ with respect to v_k.

Now the gradient vector (4.19) and the observed information matrix (4.20) can be calculated using the above quantities. Estimates of θ are updated using the Newton-Raphson method,

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} + \left[\frac{-\partial^2 h_A}{\partial \theta^2}\right]^{-1} \frac{\partial h_A}{\partial \theta}\bigg|_{\theta = \hat{\theta}^{(i)}} \tag{4.34}$$

where $\hat{\theta}^{(i)}$ is the estimate of θ at the ith iteration.

In summary, the MHLE of β, v and θ are found by using the Newton-Raphson method to maximize the profile h-likelihood and the adjusted profile h-likelihood. First, the estimates of β and v are updated with a single Newton-Raphson step on the profile h-likelihood conditional on the current estimate of θ. Then the estimate of θ is updated with a single step of Netwon-Raphson to maximize the adjusted profile h-likelihood, given the current estimates of β and v. Continue alternating between (4.9) and (4.34) until convergence is achieved. Convergence is defined as,

$$\max\left\{\left|\hat{\beta}^{(i+1)} - \hat{\beta}^{(i)}\right|, \left|\hat{\theta}^{(i+1)} - \hat{\theta}^{(i)}\right|\right\} < \Delta,$$

where Δ is a predetermined tolerance limit. After convergence has been achieved, the estimated covariance matrix of the parameter estimates is evaluated by taking the inverse of the observed information matrix (4.15) for the profile h-likelihood.

4.2 ESTIMATING THE CAUSE-SPECIFIC HAZARD UNIVARIATE FRAILTY MODEL

Assuming a multivariate distribution for the random effects allows for a more general model. However assuming a univariate distribution requires less computation and is easier to implement. Fitting the cause-specific hazard univariate frailty model is very similar to the multivariate case. This section outlines the main modifications to the method presented in the previous section for the multivariate case.

The h-likelihood assuming a univariate normal distribution is,

$$h(\beta, \lambda_0, v, \theta) = \sum_{ijk} h_{ijk} = \sum_{ijk} l_{ijk}(\beta_k, \lambda_{0k}; t_{ij}, \delta_{ijk}|v_i) + \sum_i l_i(\theta; v_i) \tag{4.35}$$

where,

$$l_{ijk}(\beta_k, \lambda_{0k}; t_{ij}, \delta_{ijk}|v_i) = \delta_{ijk}\left(\log \lambda_{0k}(t_{ij}) + x_{ij}^T\beta_k + v_i\right) - \Lambda_{0k}(t_{ij}) \exp\left(x_{ij}^T\beta_k + v_i\right)$$

is the log of the conditional density function for T_{ij} and δ_{ijk} given $V_i = v_i$ and

$$l_i(\theta; v_i) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\theta) + \frac{-v_i^2}{2\theta} \tag{4.36}$$

is the log of the univariate normal density function with mean 0 and variance θ. Since the random effects no longer depend on the event type V_i represents the effect for any event type in cluster i and $V = (V_1, V_2, \ldots, V_n)$ is a n-dimensional vector of all the random effects for every cluster.

The profile h-likelihood is very similar to (4.8),

$$h_p(\beta, v, \theta) = \sum_{kr} S_{Xkr}^T\beta_k + S_{Zkr}^T v - d_{kr}\log\left(\sum_{ij\in\mathcal{R}_{kr}} \exp\left(X_{ij}^T\beta_k + Z_{ij}^T v\right)\right) - d_{kr} + \sum_{i=1}^n l_i(\theta; v_i), \tag{4.37}$$

where the only change is to the treatment of the random effects. Differentiating (4.37) with respect to β is identical to (4.12), while the $\partial h_p/\partial v$ is different,

$$\frac{\partial h_p}{\partial v} = \sum_k Z^T(E_k - M_k) - \frac{v}{\theta}. \tag{4.38}$$

Likewise, there are also changes to the observed information matrix, which is smaller and no longer block diagonal. Now the observed information of β and v is a $(mp + n) \times (mp + n)$ matrix H given by,

$$H = H(\beta, v, \theta) = \begin{pmatrix} X^TW_1X & \mathbf{0} & X^TW_1Z \\ \mathbf{0} & X^TW_2X & X^TW_2Z \\ Z^TW_1X & Z^TW_2X & \sum_k Z^TW_kZ + Q \end{pmatrix} \tag{4.39}$$

where $W_k = W_k(\beta_k, v_k) = M_k - N_kR_kC_k(R_kN_k)^T$ for $k = 1, 2$ and Q is a $n \times n$ diagonal matrix with ith element $-\partial^2 l_i/\partial v_i^2 = 1/\theta$. Similar modifications are necessary when taking the first and second derivatives of the adjusted profile h-likelihood.

When there is only one frailty parameter, it is possible to do higher order approximations; these higher order approximations are also possible with multiple frailty parameters but are much more difficult. Estimates of θ can be updated using either the first-order or second-order Laplace approximation (Lee and Nelder, 2001). For the shared frailty model, Ha and Lee (2003) recommend the first-order Laplace approximation for models with a lognormal frailty and the second-order Laplace approximation for models with a gamma frailty distribution. The first-order Laplace approximation was used in section 4.1 and is the same as the adjusted profile h-likelihood. The second-order Laplace approximation is defined by (Lee and Nelder, 2001),

$$h_A^\star = [h_A - \text{trace}(S)/24]\Big|_{(\beta, v) = (\hat{\beta}, \hat{v})} \tag{4.40}$$

where S is a $n \times n$ diagonal matrix where the ith element is,

$$S_i = \left[3\frac{-\partial^4 h_p/\partial v_i^4}{(-\partial^2 h_p/\partial v_i^2)^2} - 5\frac{(-\partial^3 h_p/\partial v_i^3)^2}{(-\partial^2 h_p/\partial v_i^2)^3}\right]_{v_i = \tilde{v}_i} \tag{4.41}$$

and $\tilde{v}_i$ is the solution to $\partial h_p/\partial v_i = 0$ given β. The first derivative of the second-order approximation is,

$$\frac{\partial h_A^\star}{\partial \theta} = \frac{\partial h_A}{\partial \theta} - \frac{1}{24}\text{trace}\left(\frac{\partial S}{\partial \theta}\right) \tag{4.42}$$

and the observed information is,

$$-\frac{\partial^2 h_A^\star}{\partial \theta^2} = -\frac{\partial^2 h_A}{\partial \theta^2} - \frac{1}{24}\text{trace}\left(-\frac{\partial^2 S}{\partial \theta^2}\right). \tag{4.43}$$

For a lognormal frailty distribution a closed form expression for $\tilde{v}_i$ does not exist. So numerical methods are used to approximate $\tilde{v}_i$. For a gamma frailty with mean 1 and variance θ a closed form solution does exist. Using a second-order Laplace approximation estimates of θ are updated by,

$$\theta^{(i+1)} = \theta^{(i)} + \frac{\partial h_A/\partial\theta - \frac{1}{24}\text{trace}(\partial S/\partial\theta)}{-\partial^2 h_A/\partial\theta^2 - \frac{1}{24}\text{trace}(\partial^2 S/\partial\theta^2)}\Bigg|_{\theta\,=\,\hat{\theta}^{(i)}}. \tag{4.44}$$

4.3 ESTIMATING THE SUBHAZARD FRAILTY MODEL

Estimation of the subhazard frailty model is very similar to estimating the cause-specific hazard univariate frailty model in section 4.2 as well as the shared frailty model. The profile h-likelihood for the subhazard frailty model is,

$$h_p(\beta, v, \theta) = \sum_r S_{Xr}^T \beta + S_{Zr}^T v - d_r \log \left(\sum_{r \in \mathcal{R}_{ij}} w_{ij}(t_r) \exp\left(X_r^T \beta + Z_r^T v\right) \right) - d_r + \sum_{i=1}^{n} l_i(\theta; v_i). \tag{4.45}$$

where the at risk set $\mathcal{R}_{ij}$ and weight function $w_{ij}(t_r)$ are defined in section 3.5 and l_i is the log of a normal density with mean 0 and variance θ.

Notice that (4.45) is similar to the profile h-likelihood for the univariate cause-specific hazard frailty model (4.37). There are a few differences, first there is a weight function $w_{ij}(t_r)$; second the at risk set is different, it includes both individuals who have not experienced an event and those who experienced a competing event; and third there is only one event type $m = 1$ so there is no subscript k.

Given these differences to the profile h-likelihood, the easiest way to modify the method in section 4.2 to fit the subhazard frailty model is to set $m = 1$ since only the event of interest is being modeled. As well as replace the at risk indicator matrix R_k with a weighted at risk indicator matrix R_k^* that contains the weights w as well as the at risk set used for modeling the subhazard function. Let $R_k^* = (R_1^*, R_2^*, \ldots, R_{D_k}^*)$ be a $N \times D_k$ at risk weighted indicator matrix where the ijth element in column r for event type k is

$$I(t_{ij} \geq t_{kr} | \delta_{ijk} \neq 1) \times \frac{\hat{G}(t_{kr})}{\hat{G}(\min(t_{kr}, t_{ij}))} \tag{4.46}$$

where $\hat{G}$ is the Kaplan-Meier estimate of the survival function for the censoring times. Since the subhazard model assumes a univariate normal distribution it is possible to do higher order approximations for this model.

4.4 INFERENCE AND MODEL SELECTION

From Lee et al. (2006), the profile h-likelihood (4.8) can be treated like an ordinary likelihood for fixed θ. This means that Wald hypothesis tests and confidence intervals can be used for approximate inference on regression coefficients β and random effects v. Let the partitioned matrix,

$$H^{-1} = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{pmatrix} \tag{4.47}$$

be the inverse of the observed information matrix H (4.15) for $(\hat{\beta}, \hat{v})$ with θ fixed, where H^{11} approximates the variance of β (Pawitan, 2001). Then an approximate $100(1-\alpha)\%$ confidence interval for β is,

$$\hat{\beta} \pm Z_{\alpha/2}(H^{11})^{-1/2} \tag{4.48}$$

where $Z_{\alpha/2}$ is the $\alpha/2$ critical value for the standard normal distribution.

The inverse of the observed information matrix for the adjusted profile h-likelihood (4.17) does a poor job of approximating the variance of the frailty parameter $\hat{\theta}$. PPL methods have a similar problem and also poorly estimate the standard error of $\hat{\theta}$ (McGilchrist, 1993, Therneau et al., 2003). However, estimating the standard error of $\hat{\theta}$ is not very important. Recall that frailty parameters θ are really variance parameters for the normal distribution with parameter space $\theta \geq 0$. Therefore, Wald hypothesis test and confidence intervals cannot be used to test the hypothesis of no cluster effect $\theta = 0$ because the null hypothesis is on the boundary of the parameter space.

Nonetheless, a likelihood ratio test can still be used to test the null hypothesis $H_0 : \theta = 0$. Let $l_1(\hat{\beta}, \hat{v}, \hat{\theta})$ be the log-likelihood of a competing risks frailty model evaluated at the MHLE and let $l_2(\hat{\beta})$ be the maximum of the log-likelihood for the corresponding submodel with no random effects. Then the likelihood ratio test statistic is,

$$\hat{\psi} = -2[l_2(\hat{\beta}) - l_1(\hat{\beta}, \hat{v}, \hat{\theta})]. \tag{4.49}$$

Under the null hypothesis this test is on the boundary of the parameter space. Therefore the asymptotic distribution of the likelihood ratio statistic ψ is no longer chi-squared. Instead, from Self and Liang (1987) the asymptotic distribution of ψ is a 50:50 mixture distribution of a χ^2_0 and a χ^2_1 distribution, where χ^2_{df} denotes a chi-square distribution with df degrees of freedom. Then the p-value for the likelihood ratio test is $\frac{1}{2}P(\chi^2_1 > \hat{\psi})$,

Following Noh et al. (2006), the adjusted profile h-likelihood h_A is used for l_1 in (4.49). This approximate likelihood is just a function of θ, where the other parameters β and v as well as the nuisance parameter λ_0 have been removed. Thus it is reasonable to use h_A to test for the frailty effect. The likelihood for the submodel l_2 depends on which full model is being tested. For the subhazard frailty model, l_2 is the log-likelihood returned by the Fine and Gray (1999) model for the hazard function of the subdistribution. Since the cause-specific hazard frailty model fits all event types jointly, the log-likelihood for this model without random effects is the complete competing risks log-likelihood (2.33). This can be found by summing the log-likelihoods returned by fitting a proportional hazards model for each event type or by modeling the cause-specific hazards jointly (Lunn and McNeil, 1995). A similar likelihood ratio test was recommended by Katsahian and Boudreau (2011) to test for clustering effects when using PPL to estimate the subhazard frailty model. There are other ways to test for clustering. Gray (1992) proposes a Wald test that tests whether all random effects are 0, $H_0 : v_1 = v_2 = \ldots = 0$.

The univariate and multivariate cause-specific hazard frailty models are non-nested models; the role of the random effect(s) is different in each model. To select which model is more appropriate Akaike's information criterion (AIC) can be used (Ha et al., 2007). Define the AIC criteria as,

$$AIC = -2h_A(\hat{\theta}) + 2s, \tag{4.50}$$

where $h_A(\hat{\theta})$ is the maximum adjusted profile h-likelihood and s is the number of variance components in Σ; s is not the total number of parameters. The model with the smaller AIC indicates a better fitting model. The term $2s$ acts like a penalty increasing the AIC for using a more complex correlation structure that will most likely increase h_A and decrease the AIC.

Using the AIC, a more complex model with more parameters will only be selected if there is a marked improvement in the model fit.

The AIC criteria can also be used to select which correlation structure for Σ gives the best fit when using the multivariate model, as well as choose between a model with random effects and without random effects.The AIC for a model without random effects is simply -2 times the log-likelihood from the corresponding competing risks model. There are no variance components so $s = 0$. However it may make more sense to fit the simpler univariate frailty model and test the cluster effect using (4.49). This AIC criteria cannot be used to select fixed effects β since they have been eliminated from h_A; the adjusted profile h-likelihood is just a function of the frailty parameters θ.

5.0 SIMULATION

Simulations by Ha et al. (2001) and Ha and Lee (2003) have demonstrated that h-likelihood performs well under a variety of scenarios when estimating the shared frailty model. Moreover simulations results by Ha et al. (2011) have also shown that h-likelihood performs well assuming a bivariate frailty distribution when fitting the shared frailty model with two random effects. All of the previously mentioned simulations have also shown that misspecification of the frailty distribution has a minimal effect on the estimated covariates when using h-likelihood.

In this chapter simulation studies are performed to evaluate the performance of the h-likelihood estimation method when fitting competing risks frailty models. First, section 5.1 demonstrates the performance of using h-likelihood to fit the cause-specific hazard frailty model assuming both a univariate and bivariate normal distribution. This section also uses simulation to evaluate the accuracy of the AIC criteria for selecting the most appropriate covariance structure for the bivariate frailty distribution from a collection of possible covariance structures. Then section 5.2 compares h-likelihood and the PPL method proposed by Katsahian and Boudreau (2011) for fitting the subhazard frailty model. Simulation results by Katsahian and Boudreau (2011) demonstrate that PPL does well when the cluster size is large. Therefore the purpose of this section is to compare h-likelihood and PPL for cases when the cluster size is small. Additional simulations examine the consequences of ignoring clustering when analyzing competing risks data. Each section includes details on how to simulate competing risks data in general. Simulating competing risks data is not as simple as using the probability integral transformation, where data is generated from a distribution function $F(x)$ by solving the equation $F(x) = u$ for x where u is a random standard uniform number.

The performance of the estimators $\hat{\theta}$ for estimating the parameters θ are evaluated and compared using the relative bias or percent bias,

$$\%\text{Bias}(\hat{\theta}) = \frac{E(\hat{\theta}) - \theta}{\theta} \times 100 \tag{5.1}$$

and the mean square error,

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2, \tag{5.2}$$

as well as the coverage probability for a confidence level of 95%, the proportion of samples whose 95% confidence interval contains the true value of the parameter.

5.1 CAUSE-SPECIFIC HAZARD FRAILTY MODEL

5.1.1 Data Generation

Event times for the cause-specific hazard frailty model are generated using a procedure similar to the one given by Beyersmann et al. (2009). This approach uses the cause-specific hazard functions rather than assume a latent failure time model. A latent failure time model (Pintilie, 2006) assumes that there is an unobserved event time for each event type. Of these hypothetical event times, only the minimum event time is observed. Simulating competing risks event times from a latent failure time model by assuming some multivariate distribution for all of the event times and selecting the minimum event time is valid. However, this method does not reflect how competing risks data is actually analyzed. It is well known that the dependence structure between latent failure times cannot be identified from the observed data where only one of the competing events is recorded for each individual (Tsiatis, 1975). So, rather than simulating data using a method that requires specifying an unidentifiable dependence structure, the following procedure only relies on identifiable quantities by using the cause-specific hazard functions.

Suppose there are two event types, Type I and Type II. Then the data generation procedure is based on the following observation; given that an individual has an event at time $T = t$ the probability that the individual fails from the Type I event is,

$$\begin{aligned} P(\delta_1 = 1 | T \in \Delta t, T \geq t) &= \frac{P(T \in \Delta t, \delta_1 = 1 | T \geq t)}{P(t \in \Delta t | T \geq t)} \\ &= \frac{\lambda_1(t)}{\lambda_1(t) + \lambda_2(t)}, \end{aligned} \tag{5.3}$$

where Δt denotes both the length of the infinitesimal interval $[t, t+\Delta t)$ as well as the interval itself; δ_1 is the Type I event indicator and $\lambda_1(t)$ and $\lambda_2(t)$ are the cause-specific hazards for the Type I and Type II event, respectively (section 2.2).

The general data generation procedure is as follows,

1. Generate appropriate covariate values X for each observation.
2. Generate n random effects v, one per cluster, from an assumed distribution.
3. Specify the cause-specific hazards $\lambda_1(t|X, v)$ and $\lambda_2(t|X, v)$ conditional on the covariates and random effect.
4. Simulate survival times T from the overall hazard function $\lambda_1(t|X, v) + \lambda_2(t|X, v)$ using the probability integral transformation (Bender et al., 2005).
5. Determine which event type is associated with the simulated time T by running a Bernoulli experiment which selects a Type I event with probability,

$$\frac{\lambda_1(t|X, v)}{\lambda_1(t|X, v) + \lambda_2(t|X, v)}.$$

6. Generate independent non-informative censoring times C.
7. Select the observed event time as the minimum of T and C and create a variable to indicate the event type and censoring status.

If there are more than two event types replace the Bernoulli experiment in step 5 with a multinomial experiment.

Using this data simulation method, data is generated for the two non-nested cause-specific hazard frailty models from section 3.4. The first is (3.35) which assumes a univariate normal distribution. The second is (3.36) which includes a random effect for each event type

and assumes a bivariate normal distribution. The simulation scheme is the same for both model. The only difference is the distribution of the random effects.

Let there be two event types, Type I and Type II as well as independent censoring. Samples sizes of $N = 100$ and $N = 200$ are considered where $(n, n_i) = (50, 2),\ (50, 4)$ and $(100, 2)$. Data were generated with two covariates (X_{ij1}, X_{ij2}), where X_{ij1} follows a standard normal distribution and X_{ij2} is a Bernoulli random variable with probability 0.5. For the univariate case the random effects V_i are generated from a $N(0, \theta)$ distribution with $\theta = 1$. For the bivariate case the random effects are bivariate normal,

$$\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right) \tag{5.4}$$

where the true values of the variance components are,

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \tag{5.5}$$

and the correlation between the two event types is $\rho = 0.5$; $\theta = (\sigma_{11}, \sigma_{22}, \sigma_{12}) = (1, 1, 0.5)$. For the bivariate case, the conditional cause-specific hazard rates for each event type are (for the univariate case replace v_{i1} and v_{i2} with v_i),

$$\begin{aligned} \lambda_{ij1}(t|x_{ij}, v_{i1}) &= 0.5 \exp(0.6x_{ij1} - 0.4x_{ij2} + v_{i1}) \\ \lambda_{ij2}(t|x_{ij}, v_{i2}) &= 2 \exp(-0.3x_{ij1} + 0.7x_{ij2} + v_{i2}) \end{aligned}$$

That is, $\beta_1 = (\beta_{11}, \beta_{12}) = (0.6, -0.4)$ and $\beta_2 = (\beta_{21}, \beta_{22}) = (-0.3, 0.7)$. Under this scenario, approximately 65% of the events are Type I and 35% are Type II when there is no censoring. The baseline hazards were selected to control the proportion of each event type. Censoring times are generated from a Uniform(0, τ) distribution where the value of τ is empirically selected to achieve the approximate right censoring rate, 0% and 30%. For each scenario, 500 datasets were generated.

5.1.2 Results

The simulation results for the univariate case where (3.35) is the true model are give in Table 5.1. This table as well as subsequent tables report the mean and standard deviation of the estimates from the 500 iterations as well as the average of the estimated standard errors for the regression coefficients. The last three columns give the percent bias, mean square error and coverage probability, respectively.

Whether there is no censoring or 30% censoring, h-likelihood returns estimates close to the true value of the parameters for the Type I effects, β_{11} and β_{12}. There is some difficulty estimating the coefficients for the Type II event when the sample size is small. This was expected since there are fewer Type II events than Type I events. Estimation of β_{21} and β_{22} is improved with larger sample sizes; there is a greater improvement when the cluster size increases. Larger cluster sizes rather than more clusters also give better estimates of the frailty parameter θ when there is censoring. Increasing the censoring rate does not seem to have a major impact on the performance of the estimates.

The estimated standard errors of the regression coefficients tend to underestimate the empirical standard deviation for most of the scenarios, but are pretty close. Similar results were shown by Ha et al. (2001) and Ha and Lee (2003) for the shared frailty model using h-likelihood. Ripatti and Palmgren (2000) also reported underestimated standard errors with the PPL estimator for the shared frailty model. This results in coverage probabilities being less than 95%. Nonetheless all of the coverage probabilities in this simulation still remain in the 92%-96% range. Larger cluster sizes tended to have larger coverage probabilities and more accurate estimated standard errors. As expected, the mean square error decreases for larger samples. There is a larger decrease in MSE when the cluster size is double compared to when the number of clusters is doubled. Ha and Lee (2003) also demonstrated that increasing the cluster size rather than the number of clusters had a bigger impact on reducing the variance and bias of the estimator for the shared frailty model.

With 0% censoring, 97% of the samples converge when $n = 50$ and $n_i = 2$ and 1 sample failed to converge when $n = 100$ and $n_i = 2$. These samples were not included in Table 5.1. For the remaining scenarios, all of the samples converged.

Table 5.1: Simulation results, cause-specific hazard frailty model - univariate case

Censoring	Sample Size	Parameter	Mean	SD	SE	%Bias	MSE	CP
0%	$n = 50,\ n_i = 2$	β_{11}	0.614	0.175	0.172	2.3	0.031	95.7
		β_{12}	-0.412	0.343	0.319	3.1	0.118	94.3
		β_{21}	-0.340	0.246	0.231	13.4	0.062	93.3
		β_{22}	0.758	0.480	0.443	8.3	0.234	93.3
		θ	1.016	0.539		1.6	0.290	
	$n = 50,\ n_i = 4$	β_{11}	0.601	0.112	0.111	0.1	0.012	95.6
		β_{12}	-0.386	0.202	0.206	-3.6	0.041	96.2
		β_{21}	-0.299	0.157	0.150	-0.5	0.024	94.2
		β_{22}	0.706	0.279	0.289	0.8	0.078	96.6
		θ	1.018	0.345		1.8	0.120	
	$n = 100,\ n_i = 2$	β_{11}	0.593	0.120	0.118	-1.2	0.014	94.0
		β_{12}	-0.399	0.233	0.220	-0.1	0.054	94.2
		β_{21}	-0.311	0.166	0.157	3.5	0.028	94.8
		β_{22}	0.722	0.292	0.301	3.2	0.086	95.2
		θ	0.940	0.348		-6.0	0.125	
30%	$n = 50,\ n_i = 2$	β_{11}	0.617	0.212	0.196	2.9	0.045	94.4
		β_{12}	-0.404	0.385	0.368	0.9	0.148	94.8
		β_{21}	-0.329	0.279	0.264	9.5	0.079	94.6
		β_{22}	-0.712	0.565	0.522	1.8	0.319	94.4
		θ	1.071	0.645		7.1	0.421	
	$n = 50,\ n_i = 4$	β_{11}	0.613	0.137	0.129	2.2	0.019	92.2
		β_{12}	-0.405	0.231	0.240	1.2	0.053	95.8
		β_{21}	-0.299	0.166	0.176	-0.1	0.027	96.0
		β_{22}	-0.692	0.355	0.345	-1.1	0.126	96.0
		θ	1.010	0.388		1.0	0.151	
	$n = 100,\ n_i = 2$	β_{11}	0.598	0.141	0.134	-0.3	0.020	91.6
		β_{12}	-0.413	0.259	0.253	3.3	0.067	95.6
		β_{21}	-0.317	0.185	0.182	5.5	0.035	94.6
		β_{22}	0.731	0.381	0.360	4.5	0.146	94.6
		θ	0.921	0.372		-7.9	0.145	

Mean and SD are the mean and standard deviation of the estimates from the 500 iterations. SE is the mean of the estimated standard errors. %Bias denotes the percent bias, MSE is the mean square error and CP is the coverage probability.

Next the simulation results for the bivariate case where the true model is (3.36) are presented in Tables 5.3 and 5.4; Table 5.3 gives the results for the estimated regression coefficients and Table 5.4 gives the results for the estimated variance components.

In Table 5.3 the results for the regression coefficients are similar to the univariate case in Table 5.1. Estimators of the Type I effects (β_{11}, β_{12}) perform well with no censoring and 30% censoring. As the sample size increases the estimators become less bias. With 30% censoring there is a little bit of difficulty estimating the Type II effects, however for larger sample sizes these estimators become less bias when censoring is present.

Like Table 5.1, the estimated standard errors of the regression coefficients tend to underestimate the empirical standard deviation. As a result most of the coverage probabilities are less than 95%. Increasing the sample sizes gives more accurate estimates of the standard errors. However, the estimated standard errors for larger samples still tend to underestimate the empirical standard deviation and have coverage probabilities less than 95%. Increasing the sample size also reduces the standard deviation of the estimates and the MSE, as expected. Unlike in Table 5.1, there is no clear advantage between increasing the number of clusters n or the cluster size n_i.

Estimates of the variance components in Table 5.4 are more biased compared to the estimates of the corresponding regression coefficients presented in Table 5.3, in particular, the estimated variance for the Type II random effects σ_{22}. This was also expected since there were few Type II events in this simulation. Increasing the number of clusters rather than the cluster size improves the estimation of this parameter. Increasing the sample size also reduces the standard deviation of all the estimators as well as the MSE. For the larger sample cases, the estimates of the variance components are not as accurate as the estimated regression coefficients for the same sample sizes. Further increasing the sample size to $N = 400$ with $n = 100$ and $n_i = 4$ reduces the bias and variance of the variance component estimators as well as the estimated regression coefficients (results not shown).

A case was also considered were the correlation between the random effects was negative and the random effects were heterogeneous $(\sigma_{11} = 0.5, \sigma_{22} = 1.5, \rho = -0.5)$; for a sample size of $N = 200$ where $n = 50, n_i = 4$ with 30% censoring. In this case, the percent bias and coverage probability for the regression coefficients were similar to the corresponding results

Table 5.2: Percent bias and coverage probabilities from fitting the cause-specific hazard model and subhazard model ignoring the random effect V_i; $V_i \sim N(0, \theta)$.

		Cause-Specific Hazard Model		Subhazard Model	
θ	Parameter	%Bias	CP	%Bias	CP
0.5	β_{11}	-9.3	91.9	-12.6	88.2
	β_{12}	-3.6	96.6	-13.0	94.2
1	β_{11}	-16.6	83.2	-21.4	76.8
	β_{12}	-8.1	92.9	-20.6	92.7
2	β_{11}	-23.8	72.7	-34.0	51.4
	β_{12}	-10.0	93.3	-38.2	87.9

%Bias denotes the percent bias and CP is the coverage probability.

in Table 5.3. The estimates of the variance components were also similar to results in Table 5.4 (results not shown). Demonstrating that h-likelihood methods work under a variety of conditions, not just nicely structured scenarios.

For 0% censoring, 98% of the samples converged when $n = 50$ and $n_i = 2$; 99% converged when $n = 50$ and $n_i = 4$. For all of the remaining scenarios all of the samples converged. Samples that failed to converge were not included in Tables 5.3 or 5.4.

Ignoring the random effects and fitting the cause-specific hazard model using the proportional hazards model where all competing events are treated as censored observations results in underestimation of the true values. The bias of these estimates increases and the coverage probability drops well below the nominal 95% confidence level as the variance of the random effects increases (Table 5.2). Similar results were seen for the shared frailty model, see Henderson and Oman (1999) and the references therein.

Overall, h-likelihood provides reasonably close estimates of the true parameters as well as reasonable 95% coverage probabilities for the cause-specific hazard frailty model assuming either a univariate or bivariate normal distribution for the random effects. The results of these simulations indicate that the estimators will perform fairly well with small samples sizes and improve in efficiency and reduce bias as the sample size increases.

Table 5.3: Simulation results for β, cause-specific hazard frailty model - bivariate case

Censoring	Sample Size	Parameter	Mean	SD	SE	%Bias	MSE	CP
0%	$n = 50,\ n_i = 2$	β_{11}	0.616	0.188	0.177	-2.7	0.036	95.1
		β_{12}	-0.391	0.342	0.331	2.2	0.117	94.7
		β_{21}	-0.307	0.248	0.231	-2.2	0.061	94.5
		β_{22}	0.701	0.457	0.443	-0.1	0.208	96.3
	$n = 50,\ n_i = 4$	β_{11}	0.602	0.117	0.115	0.3	0.014	93.0
		β_{12}	-0.418	0.219	0.214	4.6	0.048	93.4
		β_{21}	-0.297	0.156	0.152	-0.9	0.024	94.6
		β_{22}	0.709	0.304	0.292	1.3	0.093	94.8
	$n = 100,\ n_i = 2$	β_{11}	0.585	0.128	0.122	-2.5	0.017	92.8
		β_{12}	-0.391	0.219	0.228	-2.3	0.048	96.4
		β_{21}	-0.302	0.167	0.157	0.6	0.028	93.4
		β_{22}	0.698	0.306	0.303	-0.2	0.094	95.6
30%	$n = 50,\ n_i = 2$	β_{11}	0.620	0.222	0.199	3.3	0.050	93.0
		β_{12}	-0.428	0.413	0.377	7.0	0.171	92.2
		β_{21}	-0.324	0.297	0.271	8.0	0.089	93.0
		β_{22}	0.780	0.560	0.534	11.5	0.320	96.1
	$n = 50,\ n_i = 4$	β_{11}	0.597	0.139	0.130	-0.5	0.019	94.2
		β_{12}	-0.406	0.260	0.247	1.5	0.067	93.6
		β_{21}	-0.295	0.188	0.176	-1.6	0.035	93.4
		β_{22}	0.732	0.364	0.348	4.6	0.133	93.6
	$n = 100,\ n_i = 2$	β_{11}	0.586	0.140	0.135	-2.4	0.020	96.2
		β_{12}	-0.391	0.263	0.257	-2.1	0.069	94.4
		β_{21}	-0.301	0.187	0.180	0.4	0.035	94.8
		β_{22}	0.668	0.351	0.353	-4.5	0.124	95.2

Mean and SD are the mean and standard deviation of the estimates from the 500 iterations. SE is the mean of the estimated standard errors. %Bias denotes the percent bias, MSE is the mean square error and CP is the coverage probability.

Table 5.4: Simulation results for θ, cause-specific hazard frailty model - bivariate case

Censoring	Sample Size	Parameter	Mean	SD	%Bias	MSE
0%	$n = 50,\ n_i = 2$	σ_{11}	1.059	0.565	-5.9	0.322
		σ_{22}	1.262	0.947	-26.2	0.966
		σ_{12}	0.478	0.577	4.4	0.333
		ρ	0.421	0.389	-15.8	0.158
	$n = 50,\ n_i = 4$	σ_{11}	0.999	0.360	-0.1	0.130
		σ_{22}	1.105	0.535	10.5	0.297
		σ_{12}	0.501	0.328	0.3	0.108
		ρ	0.487	0.250	-2.6	0.063
	$n = 100,\ n_i = 2$	σ_{11}	0.969	0.390	-3.1	0.153
		σ_{22}	1.001	0.458	0.1	0.210
		σ_{12}	0.440	0.339	-12.0	0.118
		ρ	0.446	0.276	-10.8	0.079
30%	$n = 50,\ n_i = 2$	σ_{11}	1.179	0.803	17.9	0.676
		σ_{22}	1.491	1.625	49.1	2.882
		σ_{12}	0.449	0.750	-10.3	0.566
		ρ	0.352	0.458	-29.6	0.232
	$n = 50,\ n_i = 4$	σ_{11}	1.013	0.429	1.3	0.184
		σ_{22}	1.107	0.550	10.7	0.314
		σ_{12}	0.503	0.365	0.5	0.133
		ρ	0.479	0.265	-4.2	0.071
	$n = 100,\ n_i = 2$	σ_{11}	0.947	0.422	-5.3	0.181
		σ_{22}	1.042	0.527	4.2	0.146
		σ_{12}	0.478	0.381	-4.4	0.279
		ρ	0.475	0.279	-5.0	0.078

Mean and SD are the mean and standard deviation of the estimates from the 500 iterations. %Bias denotes the percent bias and MSE is the mean square error.

Table 5.5: Percent of samples selected using AIC for each model versus the true model; diagonal elements give the percentage of samples that were correctly selected.

	AIC Selected Model		
True Model	M1	M2	M3
M1	**93.4**	2.2	4.4
M2	5.0	**80.8**	14.2
M3	38.8	42.4	**18.8**

M1 = Cause-specific hazard frailty model with one random effect
M2 = Cause-specific hazard frailty model with two independent random effects
M3 = Cause-specific hazard frailty model with two correlated random effects

Table 5.5 gives the results of a simulation study investigating the use of AIC to select the most appropriate model. A sample of size $N = 200$ with $n = 50$ and $n_i = 4$ and 30% censoring were used. Baseline hazard rates and covariates are the same as the previous simulations. Three structures for the random effects were used,

(M1) Cause-specific hazard frailty model with one random effect (3.35) that follows a standard normal distribution,

$$v_i \sim N(0, 1)$$

(M2) Cause-specific hazard frailty model with two independent random effects (3.36) that follows a bivariate normal distribution,

$$\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

(M3) Cause-specific hazard frailty model with two correlated random effects (3.36) that follows a bivariate normal distribution,

$$\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

Notice that (M1) only has one variance component, (M2) has two variance component and (M3) has three variance components.

From Table 5.5, AIC does a good job selecting the correct model when the true model is (M1) or (M2); 93.4% of samples are correctly selected when (M1) is the true model and 80.8% of samples when (M2) is the true model. However there is difficulty selecting the correct model when the true model is (M3), only 18.8% of the samples were correctly selected. This is most likely because the AIC for (M3) has a larger penalty term since three variance components are being estimated. The average difference between (M2) and (M3) is just 1.16 this difference is most likely due to the larger penalty term for (M3) that is the result of estimating the extra covariance parameter. The average AIC is smallest for (M3), 1230.7 and larger for (M2), 1231.9.

For larger cluster sizes AIC has an easier time selecting the correct model (Ha et al., 2007). Increasing the sample size to $n = 50$ and $n_i = 7$ results in AIC correctly selecting (M3) 69.8% of the time as the better fitting model when this is in fact the true model. Under this scenario the incorrect models had some trouble converging: for (M1) 80% converged and for (M2) 97% converged; suggesting that convergence trouble may indicate that the model is not complex enough to support the data.

Recall from section 3.4, that for (M1) the correlation between event times cannot be negative. If the true model is (M3) but now with negative correlation $\rho = -0.5$ then (M1) is only selected 0.8% of the time while (M2) is selected 48% and (M3) is selected 51% of the time. Thus there is still trouble distinguishing between (M2) and (M3), but the model that does not allow negative association is barely selected at all.

5.2 SUBHAZARD FRAILTY MODEL

5.2.1 Data Generation

Since the effect of a covariate on the cause-specific hazard function can be very different from the effect of the same covariate on the subhazard function (Fine and Gray, 1999, Gray,

1988), a different technique needs to be used to generate event times for the subhazard frailty model. Suppose there are two event types, Type I and II where Type I is the event of interest. Following Fine and Gray (1999), think of the subhazard function for the Type I event $\gamma_1(t)$ as the hazard function for the improper random variable,

$$T^* = I(\delta_1 = 1) \times T + I(\delta_1 = 0)) \times \infty \tag{5.6}$$

where δ_1 is the event indicator for Type I events.

Let $p = P(\delta_1 = 1) = 1 - P(\delta_1 = 0)$ such that for all $t \in [0, \infty)$, $P(T^* < t|\delta_1 = 1) = F(t)$ and $P(T^* < t|\delta_1 = 0) = 0$, where $F(t)$ is a proper distribution function for T. Then the distribution function of T^* is a mixture of $F(t)$ and a degenerate random variable with a point mass of $1 - p$ at $t = \infty$. For $t \in [0, \infty)$ the distribution function of T^* is,

$$\begin{aligned} P(T^* < t) &= pP(T < t|\delta_1 = 1) + (1 - p)P(T < t|\delta_1 = 0) \\ &= pF(t) \\ &= F_1(t) \end{aligned}$$

where $F_1(t)$ is a subdistribution function for the Type I event. Similarly, when $t = \infty$,

$$\begin{aligned} P(T^* = \infty) &= pP(T = \infty|\delta_1 = 1) + (1 - p)P(T = \infty|\delta_1 = 0) \\ &= p(1 - F(\infty)) + (1 - p) \\ &= 1 - p \end{aligned}$$

The distribution function of T^* can be written as,

$$F_{T^*}(t) = P(T^* < t) = \begin{cases} F_1(t) & \text{if } t < \infty \\ 1 - p & \text{if } t = \infty \end{cases} \tag{5.7}$$

The conditional subhazard function for a Type I event for the jth observation of cluster i is,

$$\gamma_{ij1}(t|X_{ij}, v_i) = \gamma_{01}(t) \exp(X_{ij}^T \beta_1 + v_i) \tag{5.8}$$

It follows that the corresponding subdistribution function for Type I failures conditional on the covariates X_{ij} and random effects v_i is,

$$\begin{aligned} F_1(t|X_{ij}, v_i) &= 1 - \exp\left(-\int_0^t \gamma_{01}(u)\exp(X_{ij}^T\beta_1 + v_i)\, du\right) \\ &= 1 - [1 - pF(t)]^{\exp(X_{ij}^T\beta_1 + v_i)}. \end{aligned} \tag{5.9}$$

The probability of a Type I event $\delta_{ij1} = 1$ for observation ij given X_{ij} and v_i is,

$$P(\delta_{ij1} = 1|X_{ij}, v_i) = \lim_{t\to\infty} F_1(t|X_{ij}, v_i) = 1 - (1-p)^{\exp(X_{ij}^T\beta_1 + v_i)}. \tag{5.10}$$

As a result, the proper distribution function of T_{ij} conditional on a Type I cause of failure as well as X_{ij} and v_i is,

$$F(t|X_{ij}, v_i, \delta_{ij1} = 1) = \frac{1 - [1 - pF(t)]^{\exp(X_{ij}^T\beta_1 + v_i)}}{1 - (1-p)^{\exp(X_{ij}^T\beta_1 + v_i)}}. \tag{5.11}$$

Type I event times are then generated from the above distribution function using the probability integral transformation, conditional on simulated values of the covariates X_{ij} and random effects v_i. Following Fine and Gray (1999) a specific form of the subhazard function for Type II events is not specified. Recall, the estimation procedure in section 4.3 only estimates the regression parameters for the event of interest, in this case Type I events. Thus the subdistribution for Type II events is simply obtained by taking, $P(\delta_{ij2} = 1|X_{ij}, v_i) = 1 - P(\delta_{ij1} = 1|X_{ij}, v_i)$ and using an exponential distribution with rate $\exp(X_{ij}^T\beta_2 + v_i)$ for $F(t|X_{ij}, v_i, \delta_{ij2} = 1)$. In other words, the proper distribution function of T_{ij} conditional on a Type II cause of failure as well as X_{ij} and v_i is,

$$F(t|X_{ij}, v_i, \delta_{ij2} = 1) = 1 - \exp\left(-\exp(X_{ij}^T\beta_2 + v_i)t\right). \tag{5.12}$$

Like before, Type II events are simulated from the above distribution using the probability integral transformation.

Lastly, independent censoring times are simulated. Then the observed event time is the minimum of the event times and the censoring times and an indicator variable is used to denote the event type and censoring status.

Beyersmann et al. (2009) provides an alternative strategy for simulating subhazard data based on the cause-specific hazard functions. Compared to the method given by Fine and Gray (1999), Beyersmann et al. (2009) approach can take more computing time.

Following Katsahian and Boudreau (2011), assume the proper distribution of T is a unit exponential distribution, $F(t) = 1 - e^{-t}$. Data were generated with two covariates $X_{ij} = (X_{ij1}, X_{ij2})$, where X_{ij1} follows a standard normal distribution and X_{ij2} is a Bernoulli random variable with probability 0.5. The random effects V_i are generated from a $N(0, \theta)$ distribution with $\theta = \{0.5, 1, 2\}$. Sample sizes were $N = 100$ and $N = 200$ with $(n, n_i) = (50, 2)$ and $(50, 4)$. Only small cluster sizes were considered because the purpose of this simulation is to compare h-likelihood and PPL when the cluster size is small. The proportion p in (5.11) and (5.12) is the the proportion of Type I events when there is no random effect and the covariates are all 0, $p = P(\delta_{ij1} = 1 | X_{ij1} = 0, X_{ij2} = 0, v_i = 0)$. Two values of p are considered 0.3 and 0.7. The true regression coefficients for the Type I events are $\beta_1 = (\beta_{11}, \beta_{12}) = (0.6, -0.4)$. and regression coefficients for the Type II event are, $\beta_2 = (\beta_{21}, \beta_{22}) = (-0.3, 0.7)$. Censoring times are generated from a Uniform(0, τ) distribution where the value of τ is empirically selected to achieve the approximate right censoring rate, 0% and 30%. For each scenario 500 datasets were generated.

In this simulation the PPL method is also used to fit the subhazard frailty model. The PPL method was performed following Katsahian and Boudreau (2011) using the `R` function `coxme` in the `coxme` package (R Development Core Team, 2010, Therneau, 2009) with counting process notation and time dependent weights like in section 4.3.

5.2.2 Results

The simulation results for the case with 0% censoring are in Table 5.6. With no censoring h-likelihood gives more accurate estimates of the regression coefficients and the frailty parameter for large values of θ, when the population is more heterogeneous. However, when $\theta = 0.5$ PPL gives more accurate estimates. Estimates for both methods are slightly more bias and slightly more variable when $p = 0.3$ compared to when $p = 0.7$. Overall, the standard deviation of the estimates are roughly the same for both methods.

Table 5.6: Simulation results, subhazard frailty model with 0% censoring

Sample				H-Likelihood				PPL			
Size	p	θ		Mean	SD	%Bias	MSE	Mean	SD	%Bias	MSE
$n = 50$	0.7	0.5	β_1	0.613	0.183	2.2	0.034	0.604	0.181	0.7	0.033
$n_i = 2$			β_2	-0.401	0.306	0.4	0.093	-0.396	0.303	-1.0	0.092
			θ	0.551	0.408	10.3	0.169	0.485	0.390	-2.9	0.152
		1	β_1	0.603	0.185	0.5	0.034	0.593	0.182	-1.1	0.033
			β_2	-0.381	0.325	-4.7	0.106	-0.375	0.319	-6.2	0.103
			θ	0.985	0.639	-1.5	0.409	0.887	0.589	-11.3	0.360
		2	β_1	0.589	0.205	-1.9	0.042	0.575	0.201	-4.1	0.041
			β_2	-0.399	0.370	-0.1	0.137	-0.391	0.362	-2.2	0.131
			θ	1.995	1.095	-0.3	1.199	1.792	0.984	-10.4	1.013
	0.3	0.5	β_1	0.620	0.229	3.3	0.053	0.612	0.226	1.9	0.051
			β_2	-0.399	0.433	-0.1	0.187	-0.395	0.428	-1.2	0.183
			θ	0.592	0.579	18.4	0.344	0.491	0.542	-1.9	0.294
		1	β_1	0.576	0.231	-3.9	0.054	0.567	0.228	-5.5	0.053
			β_2	-0.413	0.443	3.3	0.197	-0.407	0.435	1.7	0.190
			θ	0.986	0.710	-1.4	0.504	0.859	0.643	-15.1	0.436
		2	β_1	0.584	0.236	-2.6	0.056	0.571	0.231	-4.8	0.054
			β_2	-0.392	0.451	-2.1	0.203	-0.382	0.439	-4.4	0.193
			θ	1.923	1.277	-3.9	1.637	1.682	1.121	-15.9	1.358

Mean and SD are the mean and standard deviation of the estimates from the 500 iterations. %Bias denotes the percent bias and MSE is the mean square error.

The results for the 30% censoring case are in Table 5.7. Like Table 5.6, h-likelihood gives less bias results when $\theta = 1, 2$ while PPL gives more accurate estimates for $\theta = 0.5$. The case where $p = 0.3$ and 30% censoring for a sample of size $n = 50$ and $n_i = 2$, is not included because this sample size was too small to guarantee reliable meaningful results. The biggest difference between the two methods is the estimation of θ. H-likelihood gives more accurate estimates of θ compared to PPL. Especially when there is not a lot of information, 30% censoring and $p = 0.3$ (Figure 5.1). This is most likely the result of h-likelihood using a higher order approximation to estimate θ. It is important to get an accurate estimate of the frailty parameter θ because underestimating the frailty parameter will result in underestimating the regression coefficients β as well (Henderson and Oman, 1999).

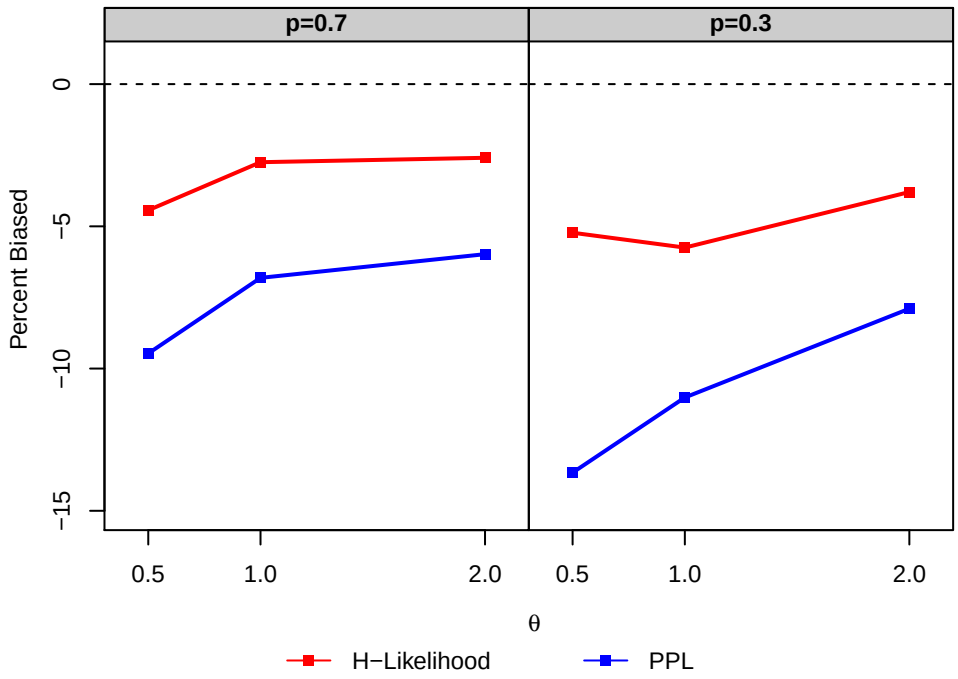

Figure 5.1: Percent bias of $\hat{\theta}$ for h-likelihood and PPL with 30% censoring and $n = 50$, $n_i = 4$

Results from Katsahian and Boudreau (2011) showed that estimates of θ overestimate the true value of the parameter. The results in Table 5.6 and 5.7 have θ being underestimated. The difference is probably due to the sample sizes. Katsahian and Boudreau (2011) used very large clusters and few clusters, while this simulation study uses small cluster sizes, but a large number of clusters. The one case Katsahian and Boudreau (2011) ran with 50 clusters of size 20 did underestimate the true value of θ.

Ignoring the frailty effect and fitting the subhazard model from Fine and Gray (1999) results in very biased estimates and low coverage probabilities (Table 5.2).

Table 5.7: Simulation results, subhazard frailty model with 30% censoring

Sample				H-Likelihood				PPL			
Size	p	θ		Mean	SD	%Bias	MSE	Mean	SD	%Bias	MSE
$n = 50$	0.7	0.5	β_1	0.601	0.198	0.1	0.592	0.592	0.198	-1.4	0.039
$n_i = 2$			β_2	-0.406	0.372	1.4	0.138	-0.399	0.370	-0.02	0.137
			θ	0.571	0.578	14.2	0.339	0.487	0.613	-2.5	0.375
		1	β_1	0.588	0.201	-1.9	0.041	0.579	0.199	-3.5	0.040
			β_2	-0.378	0.358	-5.5	0.129	-0.372	0.352	-7.0	0.125
			θ	0.909	0.619	-9.1	0.391	0.796	0.397	-20.4	0.370
		2	β_1	0.587	0.219	-2.2	0.048	0.573	0.214	-4.5	0.046
			β_2	-0.387	0.407	-3.3	0.166	-0.379	0.397	-5.3	0.158
			θ	1.825	1.105	-8.8	1.252	1.617	0.978	-19.1	1.103
$n = 50$	0.7	0.5	β_1	0.607	0.135	1.1	0.018	0.604	0.134	0.7	0.018
$n_i = 4$			β_2	-0.415	0.249	3.8	0.062	-0.413	0.248	3.3	0.061
			θ	0.478	0.276	-4.4	0.076	0.453	0.270	-9.5	0.075
		1	β_1	0.602	0.131	0.3	0.017	0.599	0.130	-0.2	0.017
			β_2	-0.389	0.248	-2.6	0.062	-0.387	0.247	-3.1	0.061
			θ	0.972	0.401	-2.7	0.161	0.932	0.384	-6.8	0.152
		2	β_1	0.596	0.130	-0.6	0.017	0.593	0.129	-1.2	0.017
			β_2	-0.417	0.239	4.2	0.058	-0.414	0.238	3.5	0.057
			θ	1.948	0.636	-2.6	0.407	1.880	0.653	-6.0	0.441
	0.3	0.5	β_1	0.594	0.175	-1.0	0.031	0.591	0.174	-1.4	0.030
			β_2	-0.418	0.341	4.5	0.117	-0.416	0.340	4.0	0.116
			θ	0.474	0.356	-5.2	0.127	0.431	0.352	-13.7	0.128
		1	β_1	0.596	0.191	-0.6	0.036	0.593	0.190	-1.2	0.036
			β_2	-0.354	0.340	-11.3	0.118	-0.353	0.339	-11.7	0.117
			θ	0.943	0.510	5.7	0.264	0.890	0.489	-11.0	0.251
		2	β_1	0.617	0.158	2.9	0.025	0.613	0.157	2.2	0.025
			β_2	-0.414	0.312	3.5	0.098	-0.411	0.310	2.7	0.097
			θ	1.92	0.840	-3.8	0.711	1.842	0.828	-7.9	0.711

Mean and SD are the mean and standard deviation of the estimates from the 500 iterations. %Bias denotes the percent bias and MSE is the mean square error.

6.0 APPLICATION

The B-14 phase III breast cancer clinical trial conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP) was a randomized double-blind multi-center trial comparing tamoxifen to placebo following surgery in patients who had negative axillary lymph nodes and estrogen receptor positive breast cancer (ER-positive). The tumors of women who are ER-positive have receptors that bind to the hormone estrogen and rely on this hormone to grow. Tamoxifen binds to estrogen receptors and inhibits the growth of the tumor by blocking estrogen from binding to the tumor.

In the B-14 trial, the placebo arm had 1413 eligible patients with follow-up and the tamoxifen arm had 1404 eligible patients with follow-up information. The study concluded that patients treated with tamoxifen had a significantly better outcome than those treated with placebo (Fisher et al., 1989, 1996). The study also observed a significant reduction of new primary cancers in the contralateral (opposite) breast for women receiving tamoxifen. This result of the B-14 study led to the P-1 Breast Cancer Prevention Trial (BCPT) where women who do not have breast cancer but are at an increased risk were randomly assigned to tamoxifen or placebo for 5 years. This study also demonstrated the benefits of tamoxifen, concluding that tamoxifen led to a 49% reduction in the risk of invasive breast cancer (Fisher et al., 1998).

In this chapter two separate analysis of the B-14 data are presented to illustrate h-likelihood procedures in real applications. First in section 6.1 the cause-specific hazard frailty model is used to estimate the effect of tamoxifen on different types of failures where some subjects experienced multiple events and competing risks are present. Next section 6.2 looks at the effect of tamoxifen on local or regional recurrence adjusting for possible center effects using the subhazard frailty model. For both analyses, p-values less than 0.05 are

considered statistically significant. Moreover all descriptives and inference for the random effects were done on the normal scale and not the lognormal scale.

6.1 REPEATED EVENTS WITH MULTIPLE TYPES OF EVENTS AND A TERMINAL EVENT

This analysis will use a high risk subset of patients from the B-14 study, those subject with a tumor size greater than 2.5 centimeters. In this subset there are 731 women (371 placebo and 360 tamoxifen) who are eligible with follow-up. The median age for women on either placebo or treatment was 55 years. The assigned therapy was administered for five years since randomization or until the first treatment failure, whichever came first. Multiple types of treatment failure are possible: local, regional, or distant recurrence of the original cancer as well as a new second primary cancer or death. Patients were followed after discontinuing therapy. Thus more than one failure type may be experienced and recorded for each subject. Since these repeated events are occurring on the same subject and each subject has a unique medical history, family history, etc it is very reasonable that the multiple event times for a patient are correlated.

The objective of this analysis is to assess the effect of treatment on local or regional recurrence and second primary cancer in the contralateral breast as well as to get an idea of the association between these two types of events. For the purpose of this analysis the types of failures will be divided into three event types. The first type (Type I) is a local or regional recurrence, the second type (Type II) is a new second primary cancer in the contralateral breast and the third type (Type III) is a distant recurrence, other new second primary cancer or death.

The occurrence of a Type III event before a Type I or Type II event can substantially change the effect of the treatment solely preventing a Type I or Type II event, because additional therapies in addition to tamoxifen may be administered. Therefore, the Type III events are competing with the Type I and Type II events; the Type I and Type II events are not competing with each other nor are they competing with the Type III event.

Table 6.1: Event type by treatment group for all observations including multiple observations from the same subject.

Event Type	Placebo	Tamoxifen	Total
Type I: Local or regional recurrence	73	40	113
Type II: Second primary in contralateral breast	32	32	64
Type III: Distant recurrence, other second primary or death	204	184	388
No events	127	148	275

Table 6.1 gives the number of events by treatment group for all observations, including multiple observations from the same subject. The most common event type was a Type III event. Subjects receiving placebo had more events for all event types, except for Type II events where both groups had exactly the same number of events. The original B-14 manuscript (Fisher et al., 1989, 1996) reported that there was a significant reduction of new primary cancers in the contralateral breast for all women receiving tamoxifen in the B-14 study. From the counts in Table 6.1 for Type II events it is clear that this analysis will not reach the same conclusion. This difference is most likely because this analysis used a subset of the original data and has less power to detect a difference between treatment groups. Among the 95 subjects who had multiple events about 57% experienced both a Type I and a Type III event and about 20% had a Type II event and a Type III event.

It is reasonable to assume that multiple event times for the same patient are correlated and that the occurrence of any one event type can affect the probability of the other event times. Thus the assumptions of the proportional hazards model are violated and this model is not appropriate. This sections fits the cause-specific hazard frailty model to the B-14 data adjusting for age and treatment assuming both a univariate and trivariate normal distribution for the random effects. This will allow estimation of the treatment effect for different types of events while accounting for their correlations and competing risks.

Table 6.2: Estimates of the cause-specific hazard frailty model univariate case; and estimates from fitting the cause-specific hazard model for each event type ignoring the effect of clustering.

		With Random Effects		Without Random Effects	
Event Type	Effect	Estimate	95% CI	Estimate	95% CI
Type I	Age	-0.015	(-0.036, 0.005)	-0.016	(-0.034, 0.002)
	Treatment	-0.742	(-1.186, -0.299)	-0.628	(-1.014, -0.243)
Type II	Age	0.002	(-0.028, 0.025)	-0.001	(-0.025, 0.024)
	Treatment	-0.179	(-0.716, 0.357)	-0.041	(-0.532, 0.449)
Type III	Age	0.016	(0.001, 0.031)	0.017	(0.007, 0.027)
	Treatment	-0.262	(-0.556, 0.032)	-0.136	(-0.336, 0.063)
Random Effect	Variance	1.895	p-value = 0.019		

CI stands for confidence interval.

The regression coefficients and estimated variance of the random effects assuming a univariate normal distribution with just one random effect per subject are given in Table 6.2. These estimates as well as the corresponding relative risks refer to comparisons within a cluster that share the same frailty. In this case clusters correspond to individual subjects. Therefore adjusting for age, the relative risk of a Type I event for an individual on tamoxifen compared to the same individual being on placebo is 0.48 with a 95% confidence interval (CI) of (0.31, 0.74). For the other event types after adjusting for age, subjects receiving tamoxifen did not have a statistically significant lower risk of experiencing the respective event type then had that same subject received placebo. A subject's age had no significant effect on any outcome. Since this is a cause-specific hazard frailty model, the estimated effects for any event type in Table 6.2 represent the pure effect of the covariate, for one type of event when the other types do not exist (Prentice et al., 1978).

The predicted cumulative incidence (Cheng et al., 1998) curves of a Type I event for a 55 years old women are given in Figure 6.1. The incidence of a Type I event increases much

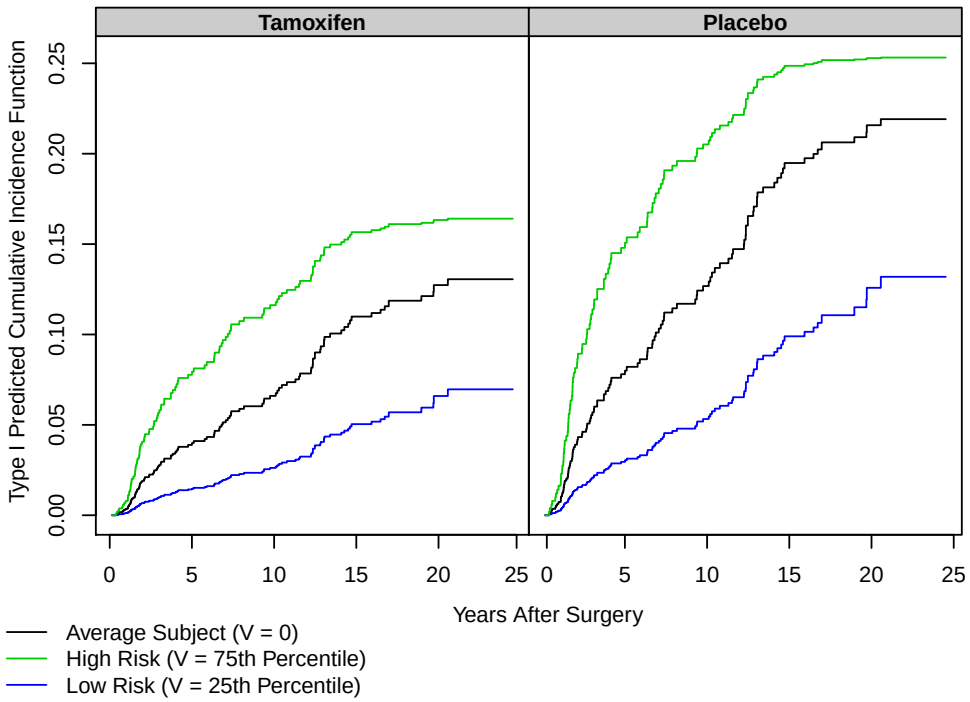

Figure 6.1: Predicted cumulative incidence of Type I events, for an average subject when $V = 0$ a high risk subject when $V = 0.82$ (75th percentile) and a low risk subject when $V = -1.07$ (25th percentile).

faster for the placebo group compared to the tamoxifen group. Ten years after surgery an average women on tamoxifen has a 7% chance of a local or regional recurrence while a women on placebo has a 13% chance. This probability increases for women who are at higher risk and decreases for women who are at low risk. The the probability of a Type I event for a high risk subject on Tamoxifen is 12% while the probability for that same women on placebo is 21%. Similarly, a low risk subject on Tamoxifen has a 3% chance and a 5% chance of a Type I event if they were on placebo.

The random effects are assumed to be normally distributed. In this case the predicted random effects are skewed right (Figure 6.2), with approximately half of the predicted effects being less than 0. The estimated 25th and 75th percentiles are respectively -1.07 and 0.82. Effects less than 0 correspond to individuals who are less frail than an average person from the study population; an average person has no effect $V = 0$. In Figure 6.2, the predicted

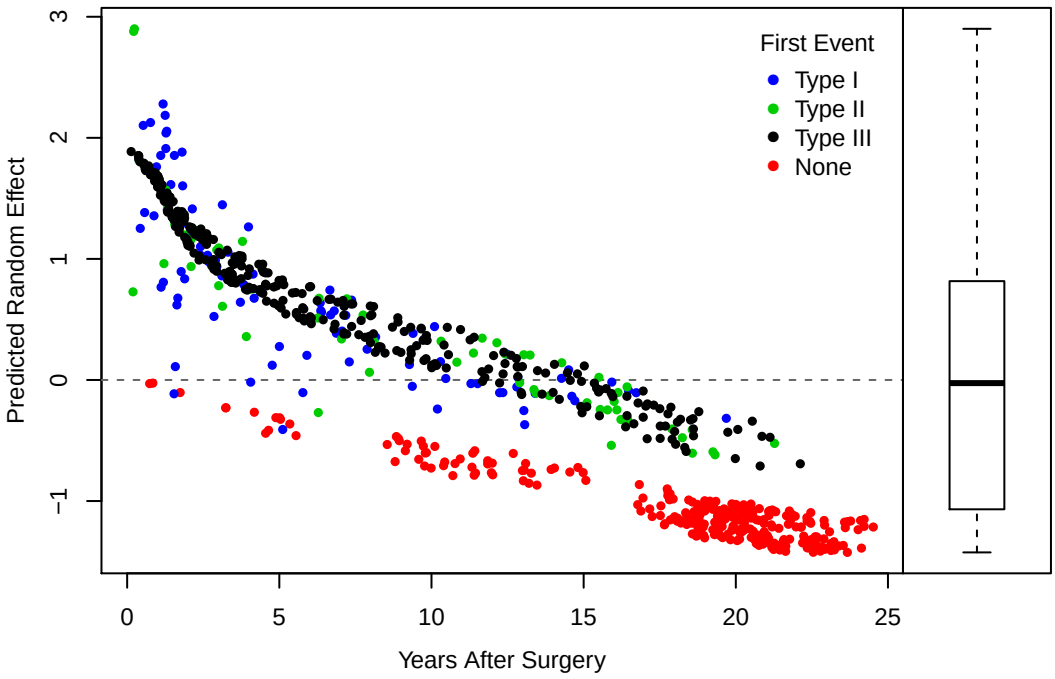

Figure 6.2: Predicted frailty versus the first observed event time for each subject; boxplot on the right hand side is the distribution of the predicted random effects.

random effects are large for subjects who had an event early and decrease for later event times. Thus those subjects who had an event early are more frail than those who survived longer (section 3.1.2). The predicted random effects for every subject who did not have an event is less than 0, this is reasonable since there is no evidence from the observed data that these subjects should be at higher risk than an average person from the study population.

The estimated variance of the random effects is 1.895 (p-value = 0.019), this suggests a fairly heterogeneous group of subjects. Moreover, the clustering effect is significant and should be considered in this analysis. Ignoring the correlation between event times and fitting the cause-specific hazard model for each event time without random effects results in much lower estimates of the treatment effect (Table 6.2). Moreover, the model with one random effect per cluster has an $AIC = 6967.8$, while the model with no random effects has an $AIC = 6970.1$. Further indicating that ignoring the correlation between event times results in a poorer fit of the data.

Table 6.3: Estimates of the cause-specific hazard frailty model trivariate case.

Event Type	Effect	Estimate	Standard Error	95% CI
Type I	Age	-0.017	0.010	(-0.036, 0.002)
	Treatment	-0.684	0.211	(-1.097, -0.271)
Type II	Age	0.002	0.013	(-0.027, 0.024)
	Treatment	-0.111	0.260	(-0.621, 0.399)
Type III	Age	0.016	0.006	(0.004, 0.028)
	Treatment	-0.203	0.125	(-0.448, 0.042)
Type I	Variance	0.789		
Type II	Variance	0.706		
Type III	Variance	0.771		
Type I, Type II	Correlation	0.886		
Type I, Type III	Correlation	0.848		
Type II, Type III	Correlation	0.897		

CI stands for confidence interval.

Additionally, the cause-specific hazard frailty model was fit assuming a trivariate normal distribution with one random effect per event type (three random effects per subject); an exchangeable correlation structure was used. This model is more general allowing negative correlation between event types, but is also more complex. Several different starting values were used to allow for the possibility of a negative association between random effects.

The estimated regression coefficients along with standard errors and confidence intervals as well as estimated variance components are given in Table 6.3. The estimated treatment effects for each event type are less than the corresponding estimates for the univariate case in Table 6.2. Subjects on tamoxifen had a significantly lower risk of a Type I event compared to subjects on placebo; tamoxifen did not significantly lower the risk for other event types. The estimated variance of the random effects for each event type are all similar ranging from 0.706 to 0.789. There is also a strong positive correlation between the random effects for

each event type, indicating that subjects who have a local or regional recurrence will also be at greater risk for a second primary cancer in the contralateral breast as well as any of the Type III events. The greater risk is because subjects who have a large random effect for a Type I event will also tend to have a large random effect for a Type II and Type III event, and larger random effects increases the risk of failure for an individual or cluster.

For this model the $AIC = 6980.2$, which is larger than the AIC for the univariate case $AIC = 6967.8$ and for the proportional hazards model $AIC = 6970.1$. Indicating that the data do not fit this model well. One possible explanation for the poor model fit is the lack of repeated observations only 95 subjects had repeated events. This model may require a larger dataset with more observations per cluster.

6.2 COMPETING RISKS WITHIN CENTERS

The purpose of this section will be to analyze the B-14 data to determine the effect of treatment on local or regional recurrence while accounting for variation between centers in the presence of competing risks. There were 167 centers in the B-14 trial. The number of subjects at each center ranged from 1 to 241, with the average center having 17 subjects and half of the centers having 8 or fewer subjects.

The analysis in this section uses all of the randomized B-14 patients; 2817 eligible patients with follow-up where 1413 were treated with tamoxifen and 1404 received placebo. The

Table 6.4: First observed event type by treatment group

Event Type	Placebo	Tamoxifen	Total
Type I: Local or regional recurrence	205	109	314
Type II: Distant recurrence, second primary or death	671	632	1303
No events	537	663	1200

Table 6.5: Estimates of the subhazard frailty model.

Event Type	Effect	Estimate	Standard Error	95% CI
Type I	Age	-0.026	0.005	(-0.037, 0.015)
	Tumor Size	0.081	0.043	(-0.003, 0.164)
	Treatment	-0.674	0.119	(-0.907, -0.441)
Center Effect	Variance	0.058		p-value = 1

CI stands for confidence interval.

average age of a subject was 55 and the average tumor size was about 2 centimeters. For this analysis, there will be two event types. The first type is, local or regional recurrence (Type I) and the second type is a new primary cancer, distant recurrence or death (Type II). Only the event that occurs first is of interest in this analysis, repeated event times are not considered. Table 6.4 gives the number of first observed event types by treatment group. Patients receiving tamoxifen experienced fewer Type I and Type II events compared to those on placebo.

The subhazard frailty model fitted the time to Type I events adjusting for age, treatment group and tumor size while accounting for the variation between centers. The results are shown in Table 6.5; recall that these effects have a center-specific interpretation. Based on this analysis, tamoxifen significantly reduced the risk of a local or regional recurrence. Women on tamoxifen had about half the risk of a Type I event compared to a women on placebo from the same center; relative risk is 0.51 with a 95% CI of (0.40, 0.64). Neither age nor tumor size were significantly associated with a Type I event.

There is no significant variation between centers, p-value=1. This lack of variation is expected in a well-designed clinical trial. In fact, the Fine and Gray (1999) model gives a better fit $AIC = 4855.0$ compared to the subhazard frailty model $AIC = 4871.5$. Since there was no significant center effect the estimates in Table 6.5 are very similar to the results returned by the Fine and Gray (1999) model (results now shown).

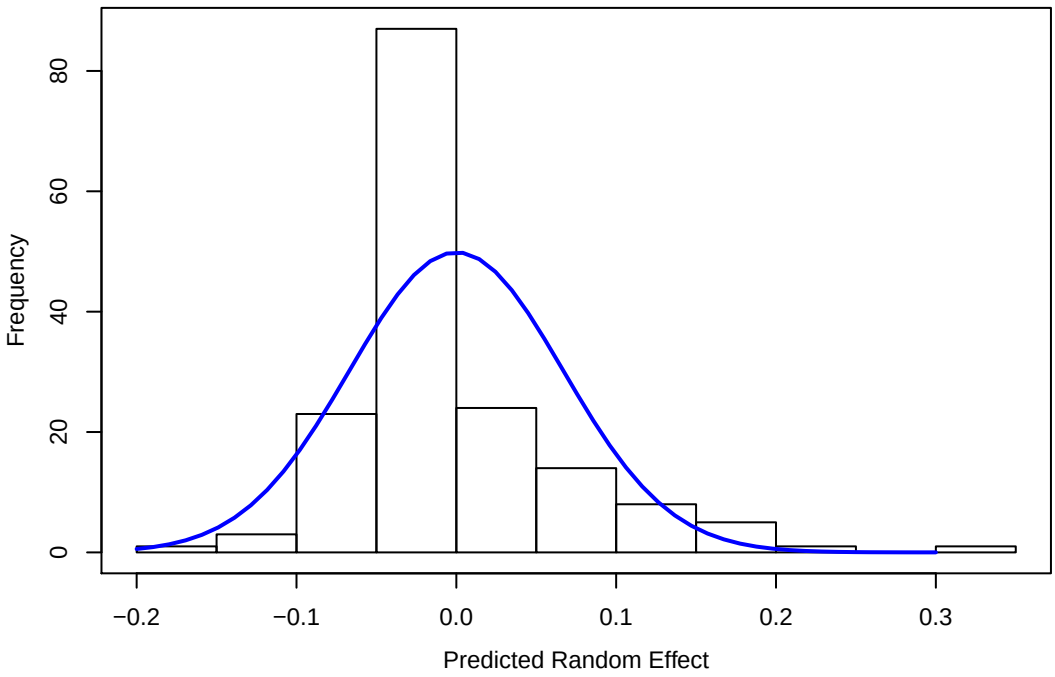

Figure 6.3: Histogram of predicted random effects with a normal density curve

Predicted random effects were very similar ranging from -0.16 to 0.3, the average was 0.04, the 25th percentile was -0.03 and the 75th percentile was 0.10. Figure 6.3 gives a histogram of the predicted frailties with an imposed normal density. The frailties are roughly symmetric with a large number of random effects just under zero.

The predicted cumulative incidence (Fine and Gray, 1999) for both treatment types is given in Figure 6.4. An average women who is 55 years old and has a tumor size of 2 centimeters has a 5% chance of a Type I event within 10 years of surgery while someone from the same center on placebo had an 9% chance. Given the little variation between centers the predicted cumulative incidence for women from a high risk center and low risk center are close to the average center.

Figure 6.5 plots the predicted cumulative incidence for an average subject of age 55 with a tumor size of 2 centimeters 20 years after surgery versus the predicted random effect for that subject's center. The solid line depicts the predicted cumulative incidence from the subhazard frailty model and the circles correspond to the observed cumulative incidences at 20 years for the cluster whose predicted random effect is given on the x-axis. The size of the

circles are proportional to the size of the center. The predicted cumulative incidence for an average women relates fairly well to the observed cumulative incidence. Centers with larger samples are close to the predicted value, while small centers with no cases are not.

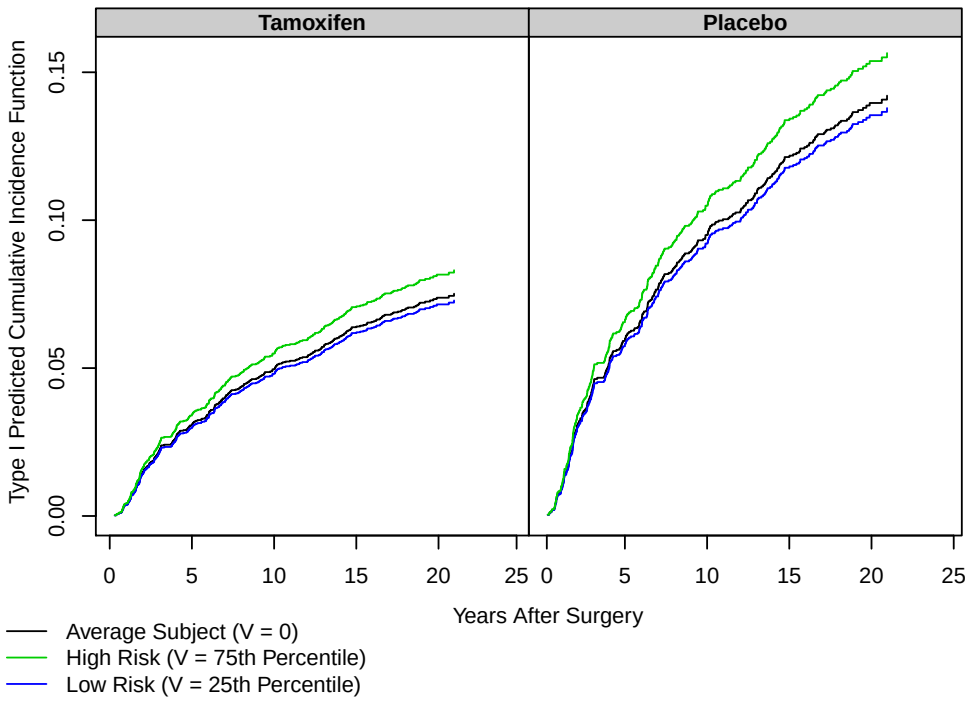

Figure 6.4: Predicted Type I cumulative incidence for subjects from: an average center $V = 0$, high risk center $V = 0.10$ (75th percentile), and low risk center $V = -0.03$ (25th percentile)

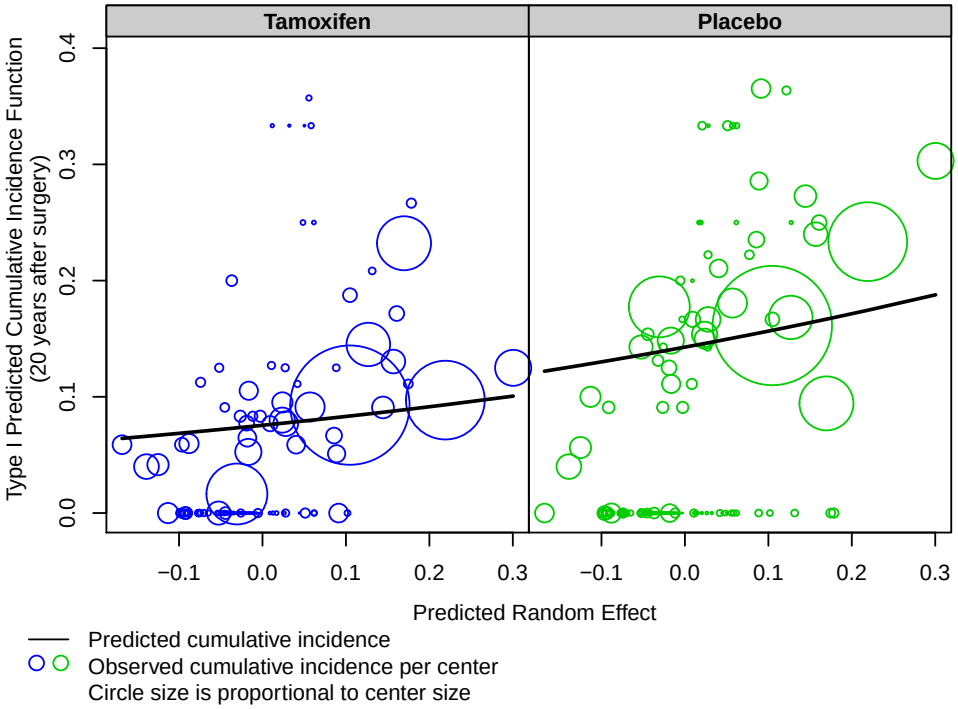

Figure 6.5: Predicted cumulative incidence versus predicted random effect.

7.0 DISCUSSION

Frailty models, an extension of the proportional hazards models, are used to model the association between clustered event times or adjust for unobserved heterogeneity by including a latent random effect that acts multiplicatively on the baseline hazard function. When competing risks are present within clusters either the cause-specific hazard frailty model or the subhazard frailty model can be used to explain the relationship between covariates and a time to event outcome.

H-likelihood provides a new estimation procedure for fitting these types of models that is computationally efficient and readily produces estimated standard errors of the parameters compared to using the EM algorithm. Furthermore, when there is little information due to high censoring and/or a rare event type, the h-likelihood will provide less biased estimates for the subhazard model compared to the PPL method; especially for the frailty parameter. Simulation results demonstrate that the h-likelihood performs well for all of the competing risks frailty models giving reasonable estimates for small cluster sizes.

A drawback of using the h-likelihood is that it can be difficult to implement because of the numerous derivatives that need to be calculated. Once the derivatives are calculated, though, the analysis is computationally efficient. Whereas the EM algorithm will always be computationally intensive. Moreover, the h-likelihood procedure presented here is mainly focused on estimating regression coefficients while accounting for the correlation between event types. This is particularly meaningful in clinical trials when the objective is to assess the effect of treatment. In other studies the correlation between event types may be more important. When a multivariate distribution is assumed it is possible to estimate the correlation between the random effects using the h-likelihood. However, there are other methods such as Bandeen-Roche and Liang (2002), Cheng et al. (2009) that maybe more appropriate

when the main objective of the analysis is the association between competing risks and not the effect of risk factors, in the presence of competing risks.

This work only considered lognormal frailty distributions. However it may be very interesting to consider other distributions when competing risks are present. In particular the compound Poisson frailty distribution. This distribution is unique in that it allows for a subgroup of zero frailty, a subgroup where no one experiences the event. While this is actually not feasible it may be a reasonable distribution when a large portion of the sample did not experience the event of interest. Consider Figure 6.2, with this example there were several people with a small predicted frailty because they did not have any events. So it would be reasonable to consider this a subgroup of people who do not experience any events. An advantage of the compound Poisson distribution is that it has a closed form Laplace transform which means that it will be simpler to calculate the marginal survival function.

Another area of future work is extending the subhazard frailty model to allow for bivariate distributed frailties. This would make the subhazard frailty model as general and flexible as the cause-specific hazard model. It would also be interesting to consider modeling individual unobserved heterogeneity in the subhazard model. Simulation results from Ha et al. (2010) show a dramatic improvement in reducing the bias of estimates in the univariate frailty model using hierarchical likelihood compared to using the marginal likelihood or penalized partial likelihood. Thus it would be worthwhile to see how the Ha et al. (2010) method applies to the subhazard frailty model.

This additional work will hopefully extend the uses of h-likelihood and provide better tools for analyzing competing risks data.

APPENDIX

R PROGRAM FOR H-LIKELIHOOD ESTIMATION

Description

Perform hierarchical likelihood estimation of the shared frailty model, cause-specific frailty model and subhazard frailty model. Assuming either a univariate normal or multivariate normal distribution for the random effects V, where different covariance structures can be assumed for the multivariate normal distribution.

Usage

```
hlike.frailty(formula, data, inits, order=1, frailty.cov="none",
        subHazard=FALSE, alpha=.05, MAX.ITER=100, TOL=1E-6)
```

Arguments

`formula`	formula, left-hand side of $\sim$ is a `CmpRsk` object (see details), right-hand side is predictors (currently limited to numeric main effects), must include a `cluster` term that identifies the cluster variable.
`data`	dataframe containing the variables used in the `formula`
`inits`	list of initial values, three named components: `beta`, `v` and `theta`
`order`	numeric, order of the Laplace approximation, 0=no order, 1=first-order, 2=second-order; second-order only applies to models with a univariate normal distribution
`frailty.cov`	character string `"none", "independent", "exchangeable"` or `"unstructured"` specifying the covariance structure for a multivariate normal distribution; `"none"` indicates univariate normal distribution
`subHazard`	logical, if `TRUE` fits the subhazard frailty model
`alpha`	numeric, 100(1-`alpha`)% confidence intervals
`MAX.ITER`	numeric, maximum number of iterations
`TOL`	numeric, tolerance limit

Details

A `CmpRsk` object is used as the response variable in the model formula. It is created using the function `CmpRsk(time, index)`, where `time` is the event time and `index` is an event indicator; values of `index` must be sequential whole numbers where 0 denotes right censoring and positive numbers refer to different event types.

If `subHazard=TRUE` then the subhazard frailty model is fit where the event of interest is when `index=1`.

Convergence is determined by,

$$\max\left\{\left|\hat{\beta}^{(i+1)}-\hat{\beta}^{(i)}\right|,\left|\hat{\theta}^{(i+1)}-\hat{\theta}^{(i)}\right|\right\}<\texttt{TOL}$$

Value

Returns a list with class `hlike` containing the following components:

`beta`	Estimated regression coefficients along with standard errors and confidence intervals
`theta`	Estimated variance components
`v`	Predicted random effects
`theta.pvalue`	P-value for testing $H_0 : \theta = 0$, only for univariate normal models
`lambda.0`	List (one component per event type) of predicted baseline hazard rate for each event time evaluated at returned estimates
`time`	List (one component per event type) of unique event times for each event type
`hp`	Profile h-likelihood evaluated at returned estimates
`Ahp`	Adjusted profile h-likelihood evaluated at returned estimates
`loglik.noFrailty`	Maximum log-likelihood for corresponding competing risks model with no frailties
`AIC`	Akaike's Information Criterion (AIC)
`converge`	Indicate convergence, 1=converge and 0=reached `MAX.ITER`
`iterations`	Number of iterations
`call`	Function call

Examples

```
# The kidney dataset is part of the survival library. It is not a competing
# risks dataset but can still serve as an example for using hlike.frailty
n = length(unique(kidney$id))
theta.init = 0.5
v.init = qnorm(runif(n, min=pnorm(-2, sd=sqrt(theta.init)),
                        max=pnorm(2, sd=sqrt(theta.init))), sd=sqrt(theta.init))
beta.init = coxph(Surv(time, status)~age+sex, data=kidney)$coef

# Shared Frailty
hlike.frailty(CmpRsk(time, status)~age+sex+cluster(id), data=kidney,
    inits=list(beta=beta.init, theta=theta.init, v=v.init),
    order=1, frailty.cov="none", MAX.ITER=300)

# Subhazard Frailty
hlike.frailty(CmpRsk(time, status2)~age+sex+cluster(id), data=kidney,
    inits=list(beta=beta.init, theta=theta.init, v=v.init),
    order=1, frailty.cov="none", subHazard=TRUE, MAX.ITER=300)

# Cause-Specific Hazard Frailty (treat censoring as a second event type)
kidney$status2=kidney$status+1

# Univariate case
beta.init <- c(sapply(1:2, function(k)
    coxph(Surv(time, status2==k)~age+sex, data=kidney)$coef))
hlike.frailty(CmpRsk(time, status2)~age+sex+cluster(id), data=kidney,
    inits=list(beta=beta.init, theta=theta.init, v=v.init),
    order=1, frailty.cov="none", MAX.ITER=100)

# Bivariate case - exchangeable
theta.init <- matrix(c(1,.5,.5,1),nrow=2)
v.init <- MASS::mvrnorm(n, mu=rep(0, 2), Sigma=theta.init)
hlike.frailty(CmpRsk(time, status2)~age+sex+cluster(id), data=kidney,
    inits=list(beta=beta.init, theta=theta.init, v=v.init),
    order=1, frailty.cov="exchangeable", MAX.ITER=100)
```

```
# hlike.frailty.R
# ------------------------------------------------------------------------------
library(survival)
library(cmprsk)
library(Matrix)

# Simple functions to assist hlike.frailty()
# ------------------------------------------------------------------------------
# Calculate the trace of a matrix
trace <- function(x) {stopifnot(is.matrix(x), nrow(x)==ncol(x)); sum(diag(x))}
# Create a block diagonal matrix
block <- function(...) {as.matrix(bdiag(...))}
# Binary operator return x without y
"%w/o%" <- function(x, y) {!x %in% y}
# Pairwise sum of multiple objects (vectors or matrices)
p.sum <- function(...) {
   x <-list(...)
   eval(parse(text=paste("x[[", 1:length(x), "]]", sep="", collapse="+")))
}

# Create an S3 class "CmpRsk"
CmpRsk <- function(time, index) {
   # Verify Arugments
   if(any(time<0)) stop("Invalid time")
   if(any(index<0, diff(sort(unique(index)))!=1)) stop("Invalid index")

   # Create object
   temp <- cbind(time, index)
   class(temp) = "CmpRsk"
   return(temp)
}

# Basic print and summary methods for a "hlike" object
print.hlike <- summary.hlike <- function(object) {
  print(object$beta); cat("\n"); print(object$theta);
  cat("\n",ifelse(object$converge==1, "Successfully Converged",
                                      "Failed to Converge"),"\n")
}
```

```
hlike.frailty <- function(formula, data, inits, order=1, frailty.cov="none",
                          subHazard=FALSE, alpha=.05, MAX.ITER=100, TOL=1E-6) {
  # Create a sorted data matrix from the given formula and dataset.
  # ------------------------------------------------------------------------
  mf <- model.frame(formula, data)
  if(!is(model.response(mf), "CmpRsk") | is.na(pmatch("cluster", names(mf))))
     stop("Invalid formula")
  resp  <- pmatch("CmpRsk", names(mf))
  clust <- pmatch("cluster", names(mf))
  pred  <- names(mf)[setdiff(1:ncol(mf), c(resp, clust))]
  temp.data <- data.frame(unclass(mf[,resp]), mf[,pred], mf[,clust])
  names(temp.data) <- c("time", "index", pred, "cluster")
  sort.data <- temp.data[order(temp.data$time),]

  N <- nrow(sort.data)                    # Total sample size
  n <- length(unique(sort.data$cluster)) # Number of clusters
  m <- max(sort.data$index)               # Number of event types
  p <- length(pred)                       # Number of predictors

  # Create design matrices for fixed covariates and the frailty term
  X <- as.matrix(sort.data[,pred]); colnames(X)=pred
  Z <- model.matrix(~ 0 + factor(sort.data$cluster)); colnames(Z)=NULL

  # Number of frailty parameters and derivative of the covariance matrix (dSigma)
  # ------------------------------------------------------------------------
  if(m==1 & frailty.cov!="none") stop("Invalid frailty.cov argument")
  if(subHazard & frailty.cov!="none") stop("Invalid frailty.cov argument")
  s <- switch(frailty.cov,
              none = 1,
              independent = m,
              exchangeable = m+1,
              unstructured = sum(1:m),
              stop("Invalid frailty.cov argument"))

  dSigma <- function(q) {
    dSigma <- matrix(0, nrow=min(m,s), ncol=min(m,s))
    zero=rep(0,s); zero[q] <- 1
    if(frailty.cov=="independent") { dSigma = diag(zero)
    } else if(frailty.cov=="exchangeable") {
         if(q<=nrow(dSigma)) dSigma <- diag(zero[-s])
         if(q>nrow(dSigma))  dSigma[lower.tri(dSigma) | upper.tri(dSigma)]=1
    } else if(frailty.cov=="unstructured") {
         dSigma[lower.tri(dSigma, diag=TRUE)] <- zero
         dSigma[upper.tri(dSigma)] <- dSigma[lower.tri(dSigma)]
    } else if(frailty.cov=="none") dSigma = 1
    return(dSigma)
  }
```

```
# If the subhazard model is being fitted (subHazard==TRUE) then the event of
# interest is event 1. The function G is the Kaplan-Meier estimate of the
# survival function for the censoring distribution.
# ------------------------------------------------------------------------------
if(subHazard) {
   m = 1
   fit.censor <- survfit(Surv(sort.data$time, sort.data$index==0)~1)
   G <- function(t) {
        # The times argument of summary.survfit requires that t be sorted.
        # After getting survival times at each t, reorder s to match the
        # original order of t. Return a vector of 1s if there is no censoring.
        s <- summary(fit.censor, times=sort(t))$surv
        if(any(sort.data$index==0)) { return(s[match(t, sort(t))])
        } else return(rep(1, length(t)))
   }
}

# Unique event times (y), number of events per unique time (d), event indicator
# (delta) and risk matrix (atRisk) for event type k, k=1,...,m
#
# The atRisk matrix is a matrix of the risk set. Columns are unique event times
# with one row for each observation. A 1 denotes that the subject is still at
# risk; with sorted data resembles a lower-triangular matrix.
# ------------------------------------------------------------------------------
# Calculate, y, d, delta, and atRisk for each event type
prepData <- function(event) {
     y <- unique(with(sort.data, time[index==event]))
     d <- with(sort.data,tapply(index[index==event],time[index==event],length))
     num.y <- length(y)
     delta <- ifelse(sort.data$index==event,1,0)

     atRisk <- matrix(0, nrow=N, ncol=num.y)
     for(j in 1:num.y) {
         if(subHazard) {
            id <- which(sort.data$time>=y[j]|sort.data$index %w/o% c(0,event))
            atRisk[id, j] <- G(y[j])/G(pmin(y[j], sort.data$time[id]))
         } else {
              id <- which(sort.data$time >= y[j])
              atRisk[id, j] <- 1
         }
     }
     return(list(y=y, d=d, delta=delta, atRisk=atRisk))
}
atRisk <- lapply(1:m, function(x) prepData(x)$atRisk)
delta  <- lapply(1:m, function(x) prepData(x)$delta)
y      <- lapply(1:m, function(x) prepData(x)$y)
d      <- lapply(1:m, function(x) prepData(x)$d)
```

```
# Variables for saving the iteration history of the parameters
beta  <- array(dim=c(p,m,MAX.ITER+1),dimnames=list(colnames(X),Type=1:m,NULL))
theta <- array(dim=c(min(m,s), min(m,s), MAX.ITER+1))
v     <- array(dim=c(n, min(m,s), MAX.ITER+1),
                dimnames=list(Group=1:n, Type=1:min(m,s), NULL))

# Initial values
if(!all(c("beta","theta","v")%in%names(inits))) stop("Invalid initial values")
if(min(m,s)!=ncol(as.matrix(inits$v))) stop("Invalid initial values")
beta[,,1] <- inits$beta
theta[,,1] <- inits$theta
v[,,1] <- inits$v

# Newton-Raphson Method
# ------------------------------------------------------------------------------
exp.eta <- lambda.0 <- Lambda.0 <- list()
diag.exp.eta <- diag.Lambda.0 <- lambda.0.sq <- list()
dexp.eta <- dlambda.0  <- dLambda.0 <- list()
ddiag.exp.eta <- ddiag.Lambda.0  <- dlambda.0.sq <- list()
d2exp.eta <- d2lambda.0 <- d2Lambda.0 <- list()
d2diag.exp.eta <- d2diag.Lambda.0 <- d2lambda.0.sq <- list()
ddiag.exp.eta<-ddiag.Lambda.0<-dlambda.0<-dlambda.0.sq<-replicate(s,list())
W   <- replicate(m, list())
dW  <- replicate(s, list())
d2W <- replicate(sum(1:s), list())

q.index = cbind(rep(1:s, times=s:1), unlist(sapply(1:s, function(q) q:s)))
colnames(q.index) = c("q1", "q2")

for(i in 1:MAX.ITER) {
    SigmaInv = solve(theta[,,i])
    SigmaInv.dSigma.SigmaInv = lapply(1:s, function(q)
       solve(theta[,,i])%*%dSigma(q)%*%solve(theta[,,i]))

    Q <- kronecker(SigmaInv, diag(n))
    Q.prime<-lapply(SigmaInv.dSigma.SigmaInv,function(x) kronecker(-x,diag(n)))
    Q.prime2 <- mapply(q1=q.index[,"q1"], q2=q.index[,"q2"], function(q1, q2)
      kronecker(SigmaInv.dSigma.SigmaInv[[q2]]%*%dSigma(q1)%*%SigmaInv +
                SigmaInv%*%dSigma(q1)%*%SigmaInv.dSigma.SigmaInv[[q2]],diag(n)),
                SIMPLIFY=FALSE)
```

```
 for(k in 1:m) {
     exp.eta[[k]]  <- as.vector(exp(X%*%beta[,k,i] + Z%*%v[,min(k,s),i]))
     lambda.0[[k]] <- as.vector(d[[k]])/as.vector(t(atRisk[[k]])%*%
                          exp.eta[[k]])
     Lambda.0[[k]] <- as.vector(atRisk[[k]]%*%diag(lambda.0[[k]])%*%
                          rep(1, ncol(atRisk[[k]])))

     diag.exp.eta[[k]]  <- diag(exp.eta[[k]])
     diag.Lambda.0[[k]] <- diag(Lambda.0[[k]])
     lambda.0.sq[[k]]   <- diag(lambda.0[[k]]^2/d[[k]])

     W[[k]] <- diag.exp.eta[[k]]%*%diag.Lambda.0[[k]] -
               diag.exp.eta[[k]]%*%atRisk[[k]]%*%lambda.0.sq[[k]]%*%
               t(atRisk[[k]])%*%diag.exp.eta[[k]]
}

# Gradient of profile h-like with respect to beta and v condtional on theta
# ---------------------------------------------------------------------------
dhp.dbeta <- unlist(lapply(1:m, function(k) t(X)%*%
              (delta[[k]]-diag.exp.eta[[k]]%*%Lambda.0[[k]])))
dhp.dv    <- do.call(ifelse(frailty.cov=="none", "p.sum", "c"),
                 lapply(1:m, function(k) t(Z)%*%(delta[[k]]-
                    diag.exp.eta[[k]]%*%Lambda.0[[k]])))
G <- c(dhp.dbeta, dhp.dv + kronecker(-SigmaInv,diag(n))%*%as.vector(v[,,i]))

# Hessian of profile h-like with respect to beta and v condtional on theta
# ---------------------------------------------------------------------------
H.XX <- block(lapply(W, function(w) t(X)%*%w%*%X))
H.XZ <- do.call(ifelse(frailty.cov=="none", "rbind", "block"),
          lapply(W, function(w) t(X)%*%w%*%Z))
H.ZX <- do.call(ifelse(frailty.cov=="none", "cbind", "block"),
          lapply(W, function(w) t(Z)%*%w%*%X))
H.ZZ <- do.call(ifelse(frailty.cov=="none", "p.sum", "block"),
          lapply(W, function(w) t(Z)%*%w%*%Z))

H <- rbind(cbind(H.XX, H.XZ),
           cbind(H.ZX, H.ZZ + Q))
H.inv <- solve(H)
```

```
  # Update estimates of beta and v
  # ------------------------------------------------------------------------
  est <- c(beta[,,i], v[,,i]) + H.inv%*%G
  beta[,,i+1] <- est[1:(m*p),1]
  v[,,i+1] <- est[-(1:(m*p)),1]

  # Gradient of the adjusted profile h-like with respect to theta given
  # beta and v
  # ------------------------------------------------------------------------
  dv.dtheta <- lapply(1:s,function(q)
   matrix(-solve(H.ZZ+Q)%*%Q.prime[[q]]%*%as.vector(v[,,i]),ncol=min(m,s)))

  if(order>0)
  for(q in 1:s) {
      for(k in 1:m) {
          dexp.eta[[k]] <- as.vector(exp.eta[[k]]*
                             (Z%*%dv.dtheta[[q]][,min(k,s)]))
          dlambda.0[[q]][[k]] <- -(d[[k]]/as.vector((t(atRisk[[k]])%*%
            exp.eta[[k]])^2))*as.vector(t(atRisk[[k]])%*%dexp.eta[[k]])
          dLambda.0[[k]] <- as.vector(atRisk[[k]]%*%
            diag(as.vector(dlambda.0[[q]][[k]]))%*%rep(1,ncol(atRisk[[k]])))

          ddiag.exp.eta[[q]][[k]]  <- diag(dexp.eta[[k]])
          ddiag.Lambda.0[[q]][[k]] <- diag(dLambda.0[[k]])
          dlambda.0.sq[[q]][[k]]   <- diag(2*lambda.0[[k]]*
                                       dlambda.0[[q]][[k]]/d[[k]])

          diag.exp.eta.atRisk = diag.exp.eta[[k]]%*%atRisk[[k]]
          dW[[q]][[k]] <- (ddiag.exp.eta[[q]][[k]]*diag.Lambda.0[[k]])+
                  (diag.exp.eta[[k]]*ddiag.Lambda.0[[q]][[k]])-
                  (ddiag.exp.eta[[q]][[k]]%*%atRisk[[k]]%*%lambda.0.sq[[k]]%*%
                      t(diag.exp.eta.atRisk))-
                  (diag.exp.eta.atRisk%*%dlambda.0.sq[[q]][[k]]%*%
                      t(diag.exp.eta.atRisk)) -
                  (diag.exp.eta.atRisk%*%lambda.0.sq[[k]]%*%
                      t(ddiag.exp.eta[[q]][[k]]%*%atRisk[[k]]))
      }
  }
if(order==0) dW<-replicate(s,list(replicate(m,list(matrix(0,nrow=N,ncol=N)))))
```

```
dH.XX <- lapply(dW,function(dW) block(lapply(dW,function(w) t(X)%*%w%*%X)))
dH.XZ <- lapply(dW, function(dW)
           do.call(ifelse(frailty.cov=="none", "rbind", "block"),
           lapply(dW, function(w) t(X)%*%w%*%Z)))
dH.ZX <- lapply(dW, function(dW)
           do.call(ifelse(frailty.cov=="none", "cbind", "block"),
           lapply(dW, function(w) t(Z)%*%w%*%X)))
dH.ZZ <- lapply(dW, function(dW)
           do.call(ifelse(frailty.cov=="none", "p.sum", "block"),
           lapply(dW, function(w) t(Z)%*%w%*%Z)))

dH.dtheta <- lapply(1:s, function(q)
                rbind(cbind(dH.XX[[q]], dH.XZ[[q]]),
                  cbind(dH.ZX[[q]], dH.ZZ[[q]] + Q.prime[[q]])))

G.theta <- sapply(1:s, function(q)
     sum(apply(as.matrix(v[,,i]), 1, function(v) -
           .5*trace(SigmaInv%*%dSigma(q))+
           .5*t(v)%*%SigmaInv.dSigma.SigmaInv[[q]]%*%v))-
           .5*trace(H.inv%*%dH.dtheta[[q]]))

# Hessian of the adjusted profile h-like with respect to theta given
# beta and v
# ------------------------------------------------------------------------
dv2.dtheta2 <- lapply(1:sum(1:s), function(q) {
   q1 = q.index[q,"q1"]
   q2 = q.index[q,"q2"]
   matrix(-solve(H.ZZ+Q)%*%((dH.ZZ[[q1]] + Q.prime[[q1]])%*%
      as.vector(dv.dtheta[[q2]]) + Q.prime[[q2]]%*%
      as.vector(dv.dtheta[[q1]])+Q.prime2[[q]]%*%
      as.vector(v[,,i])), ncol=min(m,s))
})

if(order>0)
for(q in 1:sum(1:s)) {
  q1=q.index[q,"q1"]
  q2=q.index[q,"q2"]
  for(k in 1:m) {
      d2exp.eta[[k]]  <- as.vector((Z%*%dv.dtheta[[q1]][,min(k,s)])*
           (Z%*%dv.dtheta[[q2]][,min(k,s)])*
            exp.eta[[k]]+(Z%*%dv2.dtheta2[[q]][,min(k,s)])*exp.eta[[k]])
```

```
      d2lambda.0[[k]] <- as.vector(-(lambda.0.sq[[k]]%*%t(atRisk[[k]])%*%
         ddiag.exp.eta[[q2]][[k]]%*%
         Z%*%dv.dtheta[[q1]][,min(k,s)] +
         lambda.0.sq[[k]]%*%t(atRisk[[k]])%*%diag.exp.eta[[k]]%*%
         Z%*%dv2.dtheta2[[q]][,min(k,s)] +
         dlambda.0.sq[[q2]][[k]]%*%t(atRisk[[k]])%*%diag.exp.eta[[k]]%*%
         Z%*%dv.dtheta[[q1]][,min(k,s)]))
      d2Lambda.0[[k]] <- as.vector(atRisk[[k]]%*%
         diag(as.vector(d2lambda.0[[k]]))%*%rep(1, ncol(atRisk[[k]])))

      d2diag.exp.eta[[k]]  <- diag(d2exp.eta[[k]])
      d2diag.Lambda.0[[k]] <- diag(d2Lambda.0[[k]])
      d2lambda.0.sq[[k]]   <- diag((2*lambda.0[[k]]*dlambda.0[[q1]][[k]]*
         dlambda.0[[q2]][[k]] + 2*lambda.0[[k]]*d2lambda.0[[k]])/d[[k]])

      diag.exp.eta.atRisk = diag.exp.eta[[k]]%*%atRisk[[k]]
      ddiag.exp.eta.q1.atRisk = ddiag.exp.eta[[q1]][[k]]%*%atRisk[[k]]
      ddiag.exp.eta.q2.atRisk = ddiag.exp.eta[[q2]][[k]]%*%atRisk[[k]]
      aa = d2diag.exp.eta[[k]]*diag.Lambda.0[[k]] +
           (ddiag.exp.eta[[q1]][[k]]*ddiag.Lambda.0[[q2]][[k]] +
           ddiag.exp.eta[[q2]][[k]]*ddiag.Lambda.0[[q1]][[k]]) +
           diag.exp.eta[[k]]*d2diag.Lambda.0[[k]]
      bb = (d2diag.exp.eta[[k]]%*%atRisk[[k]]%*%lambda.0.sq[[k]]%*%
             t(diag.exp.eta.atRisk)) +
           (ddiag.exp.eta.q1.atRisk%*%dlambda.0.sq[[q2]][[k]]%*%
             t(diag.exp.eta.atRisk)) +
           (ddiag.exp.eta.q1.atRisk%*%lambda.0.sq[[k]]%*%
             t(ddiag.exp.eta.q2.atRisk))
      cc = (ddiag.exp.eta.q2.atRisk%*%dlambda.0.sq[[q1]][[k]]%*%
             t(diag.exp.eta.atRisk)) +
           (diag.exp.eta.atRisk%*%d2lambda.0.sq[[k]]%*%
             t(diag.exp.eta.atRisk)) +
           (diag.exp.eta.atRisk%*%dlambda.0.sq[[q1]][[k]]%*%
             t(ddiag.exp.eta.q2.atRisk))
      dd = (ddiag.exp.eta.q2.atRisk%*%lambda.0.sq[[k]]%*%
             t(ddiag.exp.eta.q1.atRisk)) +
           (diag.exp.eta.atRisk%*%dlambda.0.sq[[q2]][[k]]%*%
             t(ddiag.exp.eta.q1.atRisk)) +
           (diag.exp.eta.atRisk%*%lambda.0.sq[[k]]%*%
             t(d2diag.exp.eta[[k]]%*%atRisk[[k]]))
      d2W[[q]][[k]] = aa-(bb+cc+dd)
   }
}
if(order==0)  d2W <- replicate(sum(1:s),
    list(replicate(m, list(matrix(0, nrow=N, ncol=N)))))
```

```
d2H.XX <- lapply(d2W, function(d2W)
            block(lapply(d2W, function(w) t(X)%*%w%*%X)))
d2H.XZ <- lapply(d2W, function(d2W)
            do.call(ifelse(frailty.cov=="none", "rbind", "block"),
            lapply(d2W, function(w) t(X)%*%w%*%Z)))
d2H.ZX <- lapply(d2W, function(d2W)
            do.call(ifelse(frailty.cov=="none", "cbind", "block"),
            lapply(d2W, function(w) t(Z)%*%w%*%X)))
d2H.ZZ <- lapply(d2W, function(d2W)
            do.call(ifelse(frailty.cov=="none", "p.sum", "block"),
            lapply(d2W, function(w) t(Z)%*%w%*%Z)))

d2H.dtheta2 <- lapply(1:sum(1:s), function(q) {
                 rbind(cbind(d2H.XX[[q]], d2H.XZ[[q]]),
                       cbind(d2H.ZX[[q]], d2H.ZZ[[q]] + Q.prime2[[q]]))})

H.theta.temp <- sapply(1:sum(1:s), function(q)
                  {q1=q.index[q,"q1"]; q2=q.index[q,"q2"]
   sum(apply(as.matrix(v[,,i]), 1, function(v) {
  .5*trace(SigmaInv%*%dSigma(q2)%*%SigmaInv%*%dSigma(q1)) +
  .5*t(v)%*%(SigmaInv%*%dSigma(q2)%*%SigmaInv.dSigma.SigmaInv[[q1]])%*%v +
  .5*t(v)%*%(SigmaInv%*%dSigma(q1)%*%SigmaInv.dSigma.SigmaInv[[q2]])%*%v}))-
  c(v[,,i])%*%Q.prime[[q1]]%*%c(dv.dtheta[[q2]]) +
  .5*trace(-H.inv%*%dH.dtheta[[q1]]%*%H.inv%*%dH.dtheta[[q2]]+
      H.inv%*%d2H.dtheta2[[q]])})

H.theta <- matrix(nrow=s, ncol=s)
H.theta[lower.tri(H.theta, diag=TRUE)] <- H.theta.temp
H.theta[upper.tri(H.theta)] <- H.theta[lower.tri(H.theta)]
```

```
# Update estimates of theta
# ------------------------------------------------------------------------
if(order==2 & s!=1) stop("Invalid order argument")
if(order==0 | order==1) {
   if(frailty.cov=="independent") {
      theta[,,i+1] <- diag(as.vector(diag(theta[,,i]) +
                        solve(H.theta)%*%G.theta))
   } else if(frailty.cov=="exchangeable") {
     temp <- c(diag(theta[,,i]),unique(theta[,,i][lower.tri(theta[,,i])]))+
           solve(H.theta)%*%G.theta
     theta[,,i+1] <- diag(temp[1:(s-1)])
     theta[,,i+1][lower.tri(theta[,,i+1])|upper.tri(theta[,,i+1])]<-temp[s]
   } else if(frailty.cov=="unstructured") {
     temp <- theta[,,i][lower.tri(theta[,,i], diag=TRUE)]+
           solve(H.theta)%*%G.theta
     theta[,,i+1][lower.tri(theta[,,i], diag=TRUE)] <- temp
     theta[,,i+1][upper.tri(theta[,,i+1], diag=TRUE)] <-
       theta[,,i+1][lower.tri(theta[,,i+1], diag=TRUE)]
   } else if(frailty.cov=="none") {
       theta[,,i+1] <- theta[,,i] + solve(H.theta)%*%G.theta
   }
 } else {
     delta.i.plus <- lapply(delta, function(x)
       tapply(x, sort.data$cluster, sum))
     mu.i.plus <- lapply(1:m, function(k)
       tapply(Lambda.0[[k]]*exp(X%*%beta[,k,i]), sort.data$cluster, sum))

     dh.dv <- function(v, j) {
       Reduce("+",delta.i.plus)[j]-exp(v)*
       Reduce("+",mu.i.plus)[j]-v/theta[,,i]}
     v.tilda <- sapply(1:n, function(x)
       uniroot(dh.dv, c(-50,50), j=x, tol=1E-3)$root)

     A <- exp(v.tilda)*Reduce("+", mu.i.plus)
     trace.dS <-sum(6*A*theta[,,i]^(-2)*(A+1/theta[,,i])^(-3)-15*
        A^2*theta[,,i]^(-2)*(A+1/theta[,,i])^(-4))
     trace.d2S<-sum(12*A*theta[,,i]^(-3)*(A+1/theta[,,i])^(-3)-18*
        A*theta[,,i]^(-4)*(A+1/theta[,,i])^(-4) -
                 30*A^2*theta[,,i]^(-3)*(A+1/theta[,,i])^(-4)+60*
                 A^2*theta[,,i]^(-4)*(A+1/theta[,,i])^(-5))

     G.theta.order2 <- G.theta - trace.dS/24
     H.theta.order2 <- H.theta - trace.d2S/24

     theta[,,i+1] <- theta[,,i] + G.theta.order2/H.theta.order2
  }
```

```
 # Convergence
 # ------------------------------------------------------------------------
 converge=0
 if(max(abs(c(beta[,,i+1], theta[,,i+1])-
            c(beta[,,i], theta[,,i])), na.rm=TRUE) < TOL){
    converge=1
    break
 }
}

 # Profile h-likelihood and Adjusted Profile h-likelihood
 # ------------------------------------------------------------------------
 h.p <- sum(sapply(1:m, function(k) sum(delta[[k]]*log(exp.eta[[k]])))) +
         - sum(sapply(1:m, function(k)
             sum(d[[k]]*log(as.vector(t(atRisk[[k]])%*%exp.eta[[k]]))))) +
           sum(apply(as.matrix(v[,,i]), 1, function(v)
             .5*log(det(SigmaInv/2/pi))-.5*t(v)%*%SigmaInv%*%v))
 A.h.p <- h.p - .5*log(det(H))+(p+n)/2*log(2*pi)
 AIC <- -2*A.h.p + 2*s  # Akaike information criterion

 # Confidence intervals and hypothesis tests
 # ------------------------------------------------------------------------
 # Estimates, standard errors, and CI for regression coefficients
 beta.SE <- sqrt(diag(as.matrix(H.inv[1:(m*p), 1:(m*p)])))
 lower.beta <- c(beta[,,i]) - qnorm(1-alpha/2)*beta.SE
 upper.beta <- c(beta[,,i]) + qnorm(1-alpha/2)*beta.SE
 beta.CI <- data.frame(rep(1:m,each=p), rep(colnames(X),times=m),
              c(beta[,,i]), beta.SE, lower.beta, upper.beta)
 colnames(beta.CI) <- c("Type", "Effect", "Estimate", "SE",
     paste(alpha/2*100, "%", sep=""), paste((1-alpha/2)*100, "%", sep=""))

 # Variance component estimates
 Var.Comp <- paste("Sigma.", unlist(sapply(1:min(m,s),
     function(q) q:min(m,s))), rep(1:min(m,s), times=min(m,s):1), sep="")
 theta.Est <- data.frame(Var.Comp,
                         Estimate=theta[,,i][lower.tri(theta[,,i], diag=TRUE)])
```

```
  # Test H0: theta=0 for cause-specific and subHazard models when univariate
  # normal distribution
  if(subHazard==TRUE) {
 loglik.noFrailty <- crr(sort.data$time, sort.data$index, X)$loglik
  } else {
    loglik.noFrailty <- sum(sapply(1:m,function(k)
       coxph(Surv(time,index==k)~X,data=sort.data)$loglik[2]))
  }
  if(frailty.cov=="none") {
     psi.hat <- -2*(loglik.noFrailty - A.h.p)
     theta.pvalue <- ifelse(psi.hat<=0,1,0.5*pchisq(psi.hat,1,lower.tail=FALSE))
  } else theta.pvalue=NULL

  # Gather and output results
  # ------------------------------------------------------------------------
  v.out <- Z%*%v[,,i]; rownames(v.out)<-sort.data$cluster;
  out <- list(beta=beta.CI, theta=theta.Est, v=v.out, theta.pvalue=theta.pvalue,
    lambda.0=lapply(1:m, function(k) matrix(lambda.0[[k]],
       dimnames=list(Time=y[[k]], Type=k))),
    time=y, hp=h.p, Ahp=A.h.p, loglik.noFrailty=loglik.noFrailty, AIC=AIC,
    converge=converge, iterations=i, call=match.call())
  class(out) <- "hlike"
  return(out)
}
```

BIBLIOGRAPHY

Aalen, O. O. (1992). Modeling heterogeneity in survival analysis by the compound poisson distribution. *The Annals of Applied Probability*, 2:951–972.

Aalen, O. O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, 3:227–243.

Bandeen-Roche, K. and Liang, K.-Y. (2002). Modelling multivariate failure time associations in the presence of competing risks. *Biometrika*, 89:299–314.

Barker, P. and Henderson, R. (2005). Small sample bias in the gamma frailty model for univariate survival. *Lifetime Data Analysis*, 11:265–284.

Barndorff-Nielsen, O. (1983). Expand+on a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70:343–365.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards model. *Statistics in Medicine*, 24:1713–1723.

Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28:956–971.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89–99.

Cheng, S., Fine, J., and Wei, L. (1998). Prediction of cumulative incidence function under the proportional hazards model. *Biometrics*, 54:219–228.

Cheng, Y., Fine, J., and Kosorok, M. (2009). Nonparametric association analysis of exchangeable clustered competing risks data. *Biometrics*, 65:385–393.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of family tendency in chronic disease incidence. *Biometrika*, 65:141–151.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B*, 34:182–220.

Cox, D. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B*, 49:1–39.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society: Series B*, 39:1–38.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer.

Fine, J. and Gray, R. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94:496–509.

Fine, J., Jiang, H., and Chappell, R. (2001). On semi-competing risks data. *Biometrika*, 88:907–919.

Fisher, B., Costantino, J., Redmond, C., and et al. (1989). A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen receptor-positive tumors. *New England Journal of Medicine*, 320:479–484.

Fisher, B., Costantino, J., Wicerham, D., Redmond, C., Kavanah, M., Cronin, W., Vogel, V., Robidoux, A., Dimitrov, N., Atkins, J., Daly, M., Wieand, S., Tan-Chiu, E., Ford, L., and Wolmark, N. (1998). Tamoxifen for prevention of breast cancer: Report of the national surgical adjuvant breast and bowel project p-1 study. *Journal of the National Cancer Institute*, 90:1371–1388.

Fisher, B., Dignam, J., Bryant, J., and et al. (1996). Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor- positive tumors. *Journal of the National Cancer Institute*, 88:1529–1542.

Garnst, A., Donohue, M., and Xu, R. (2009). Asymptotic properties and empirical evaluation of the npmle in the proportional hazards mixed-effects model. *Statistica Sinica*, 19:997–1011.

Glidden, D. and Vittinghhoff, E. (2004). Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*, 23:369–388.

Gooley, T., Leisenring, W., Crowley, J., and Storer, B. (1999). Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. *Statistics in Medicine*, 18.

Gray, R. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16:1141–1154.

Gray, R. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognoses. *Journal of the American Statistical Association*, 87:942–951.

Ha, I. and Lee, Y. (2003). Estimating frailty models via poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*, 12:663–681.

Ha, I., Lee, Y., and MacKenzie, G. (2007). Model selection for multi-component frailty models. *Statistics in Medicine*, 26:4790–4807.

Ha, I., Lee, Y., and Song, J.-K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, 88:233–243.

Ha, I., Noh, M., and Lee, Y. (2010). Bias reduction of likelihood estimators in semiparametric frailty models. *Scandinavian Journal of Statistics*, 37:307–320.

Ha, I., Sylvester, R., Legrand, C., and MacKenzie, G. (2011). Frailty modelling for survival data from multi-centre clinical trials. *Statistics in Medicine*, 30:NA.

Henderson, R. and Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society: Series B*, 61:367–379.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71:75–83.

Huang, X. and Wolfe, R. (2002). A frailty model for informative censoring. *Biometrics*, 58:510–520.

Johansen, S. (1983). An extension of cox's regression model. *International Statistical Review*, 51:165–174.

Kalbfeisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, 2nd edition.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Katsahian, S. and Boudreau, C. (2011). Estimating and testing for center effects in competing risks.

Katsahian, S., Resche-Rigon, M., Chevret, S., and Porcher, R. (2006). Analysing multicentre competing risks data with a mixed proportional hazards model for the subdistribution. *Statistics in Medicine*, 25:4267–4278.

Klein, J. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 48:795–806.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2nd edition.

Korn, E. and Dorey, F. (1992). Applications of crude incidence curves. *Statistics in Medicine*, 11:813–829.

Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society: Series B*, 58:619–678.

Lee, Y. and Nelder, J. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88:987–1006.

Lee, Y., Nelder, J., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*. Chapman & Hall.

Liu, L. and Huang, X. (2008). The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Statistics in Medicine*, 27:2665–2683.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 44:226–233.

Lunn, M. and McNeil, D. (1995). Applying cox regression to competing risks. *Biometrics*, 51:524–532.

McGilchrist, C. (1993). Reml estimation for survival models with frailty. *Biometrics*, 49:221–225.

McGilchrist, C. and Aisbett, C. (1991). Regression with frailty in survival analysis. *Biometrics*, 47:461–466.

Moeschberger, M. and Klein, J. (1995). Statistical methods for dependent competing risks. *Lifetime Data Analysis*, 1:195–204.

Murphy, S. and van der Vaart, A. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95:449–465.

Nielsen, G., Gill, R., Anderson, P., and Sørenson, T. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19:25–43.

Noh, M., Ha, I., and Lee, Y. (2006). Dispersion frailty models and hglms. *Statistics in Medicine*, 25:1341–1354.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modeling and Inference using Likelihood*. Oxford.

Pepe, M. and Mori, M. (1993). Kaplan-meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine*, 12:737–751.

Pintilie, M. (2006). *Competing Risks: A Practical Perspective*. Wiley.

Prentice, R., Kalbfleisch, J., Jr., A. P., Flournoy, N., and Farewell, V. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34:541–554.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ripatti, S., Larsen, K., and Palmgren, J. (2002). Maximum likelihood inference for multivariate frailty models using an automated monte carlo em algorithm. *Lifetime Data Analysis*, 8:349–360.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56:1016–1022.

Searle, S., Casella, G., and McCulloch, C. (1992). *Variance Components*. Wiley.

Self, S. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610.

Severini, T. (2000). *Likelihood Methods in Statistics*. Oxford.

Shih, J. and Louis, T. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51:1384–1399.

Tanner, M. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer, 3rd edition.

Therneau, T. (2009). *coxme: Mixed Effects Cox Models.* R package version 2.0.

Therneau, T., Grambsch, P., and Pankratz, V. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12:156–175.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences U.S.A.*, 72:20–22.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, 19:3309–3324.

Vaupel, J., Manton, K., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454.

Vaupel, J. and Yashin, A. (1985). Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *The American Statistician*, 39:176–185.

Wienke, A. (2011). *Frailty Models in Survival Analysis*. Chapman & Hall.

Wintrebert, C., Putter, H., Zwinderman, A., and van Houwelingen, J. (2004). Centre-effect on survival after bone marrow transplantation: Application of time-dependent frailty models. *Biometrical Journal*, 46:512–525.

Wu, L. (2010). *Mixed Effects Models for Complete Data*. CRS Press.

Xue, X. and Brookmeyer, R. (1996). Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Analysis*, 2.

Yau, K. and McGilchrist, C. (1998). Ml and reml estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine*, 17:1201–1213.

Zheng, M. and Klein, J. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82:127–138.